Εθνικό Μετσόβιο Πολυτεχνείο

Σχολη Ηλεκτρολογων Μηχανικων
Και Μηχανικων Υπολογιστων

Τομεας Τεχνολογιας Πληροφορικης &
Υπολογιστων

ΥΒΡΙΔΙΚΕΣ ΕΥΦΥΕΙΣ ΤΕΧΝΙΚΕΣ ΑΝΑΛΥΣΗΣ ΓΙΑ ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΔΕΔΟΜΕΝΑ ΥΨΗΛΩΝ ΔΙΑΣΤΑΣΕΩΝ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

του

ΧΡΗΣΤΟΥ Α. ΠΑΤΕΡΙΤΣΑ

Διπλωματούχου Ηλεκτρολόγου Μηχανικού &
Μηχανικού Υπολογιστών Ε.Μ.Π. (2001)

Αθήνα, Μάιος 2007

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων & Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης & Υπολογιστων

Πρακτικο Εξετασης Διδακτορικης Διατριβης

ΥΒΡΙΔΙΚΕΣ ΕΥΦΥΕΙΣ ΤΕΧΝΙΚΕΣ ΑΝΑΛΥΣΗΣ ΓΙΑ ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΔΕΔΟΜΕΝΑ ΥΨΗΛΩΝ ΔΙΑΣΤΑΣΕΩΝ

ΧΡΗΣΤΟΣ Α. ΠΑΤΕΡΙΤΣΑΣ

Διπλωματούχος Ηλεκτρολόγος Μηχανικός &
Μηχανικός Υπολογιστών Ε.Μ.Π. (2001)

Συμβουλευτική Επιτροπή: Ανδρέας-Γεώργιος Σταφυλοπάτης
Στέφανος Κόλλιας
Παναγιώτης Τσανάκας

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή την 2η Μαΐου 2007.

...
Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

...
Στέφανος Κόλλιας
Καθηγητής Ε.Μ.Π.

...
Παναγιώτης Τσανάκας
Καθηγητής Ε.Μ.Π.

...
Γεώργιος Παπακωνσταντίνου
Καθηγητής Ε.Μ.Π.

...
Τιμολέων Σελλής
Καθηγητής Ε.Μ.Π.

...
Κωνσταντίνα Νικήτα
Καθηγήτρια Ε.Μ.Π.

...
Μιχαήλ Βαζιργιάννης
Αναπληρωτής Καθηγητής
Οικονομικού Πανεπιστημίου
Αθηνών

Αθήνα, Μάιος 2007

Η παρούσα διδακτορική διατριβή αποτελεί υποέργο του προγράμματος: «Ηράκλειτος: Υποτροφίες έρευνας με προτεραιότητα στην βασική έρευνα».
Το Πρόγραμμα «ΗΡΑΚΛΕΙΤΟΣ» συγχρηματοδοτείται από το Ευρωπαϊκό Κοινοτικό Ταμείο (75%) και από Εθνικούς Πόρους (25%).
The Project "HRAKLEITOS" is co-funded by the European Social Fund (75%) and National Resources (25%).

ΧΡΗΣΤΟΣ Α. ΠΑΤΕΡΙΤΣΑΣ
Διδάκτωρ Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περιεχόμενα

Σχήματα

Πίνακες

ΠΡΟΛΟΓΟΣ

Η διατριβή αυτή αποτελεί το αποτέλεσμα της ερευνητικής δραστηριότητάς μου στον χώρο της υπολογιστικής νοημοσύνης και μηχανικής μάθησης από τον Νοέμβριο του 2001 έως και σήμερα. Η έρευνα αυτή πραγματοποιήθηκε στο εργαστήριο Ευφυών Υπολογιστικών Συστημάτων υπό την επίβλεψη του καθηγητή κ. Ανδρέα-Γεώργιου Σταφυλοπάτη, τον οποίο θα ήθελα να ευχαριστήσω θερμά για την βοήθεια και υποστήριξή του τόσο σε επιστημονικό όσο και προσωπικό επίπεδο αλλά και για την εμπιστοσύνη που έδειξε στο πρόσωπό μου. Η συνεισφορά του στη παρούσα εργασία υπήρξε καταλυτική και πολύτιμη.

Ευχαριστώ επίσης τα μέλη της συμβουλευτικής επιτροπής, τους καθηγητές κ. Στ. Κόλλια και κ. Π. Τσανάκα για το ενδιαφέρον και τη στήριξή τους. Οφείλω να εκφράσω τις ευχαριστίες μου στα μέλη της εξεταστικής επιτροπής κ.κ. Γ. Παπακωνσταντίνου, Τ. Σελλή, Κ. Νικήτα, καθηγητές Ε.Μ.Π., και Μ. Βαζιργιάννη, αναπληρωτή καθηγητή του Οικονομικού Πανεπιστημίου Αθηνών.

Ιδιαίτερες ευχαριστίες οφείλω και στους διδάκτορες Ε.Μ.Π. Δ. Φροσυνιώτη και Μ. Περτσελάκη, οι οποίοι εκτός από φίλοι υπήρξαν πολύτιμοι συνεργάτες και συνοδοιπόροι στην προσπάθεια αυτή. Να ευχαριστήσω ακόμα τους φίλους και συναδέλφους Α. Ραουζαίου, Α. Δροσόπουλο, διδάκτορες Ε.Μ.Π., τους Σ. Βρεττό, Χ. Φερλέ, Α. Χορταρά, Χ. Χριστάκου και Γ. Καριδάκη, υποψήφιους διδάκτορες Ε.Μ.Π., καθώς και τον Σ. Μοδέ, διπλωματούχο μηχανικό Ε.Μ.Π. για την αγαστή συνεργασία καθώς και την εποικοδομητική ανταλλαγή απόψεων και ιδεών που είχαμε όλα αυτά τα χρόνια.

Επιθυμώ επίσης να ευχαριστήσω την οικογένειά μου για την στήριξη και εμπιστοσύνη τους. Τέλος, θα ήθελα να εκφράσω ιδιαίτερες και θερμές ευχαριστίες στην σύζυγό μου Ελένη Νταουντάκη, η οποία με την στήριξη, επιμονή και υπομονή της με οδήγησε στην μακρόχρονη αυτή πορεία. Η συμβολή της στην ολοκλήρωση αυτής της εργασίας είναι αναμφίβολα ανεκτίμητη.

Χρήστος Πατερίτσας
Αθήνα, Μάιος 2007

ΠΕΡΙΛΗΨΗ

Η ανάπτυξη υβριδικών συστημάτων στον χώρο της τεχνητής νοημοσύνης και της μηχανικής μάθησης βρίσκεται τα τελευταία χρόνια ανάμεσα στα πιο δημοφιλή πεδία έρευνας. Η συνεισφορά της παρούσας διατριβής εντάσσεται αφενός στην περιοχή των υβριδικών συστημάτων με την ανάπτυξη νέων μοντέλων και αλγορίθμων μάθησης και αφετέρου στην περιοχή της εξόρυξης γνώσης από δεδομένα.

Πιο συγκεκριμένα, αρχικά αναπτύχθηκαν δυο μεθοδολογίες ομαδοποίησης με κοινό στοιχείο και στις δύο να αποτελεί η χρήση αρχικά ενός αλγόριθμου ομαδοποίησης για την δημιουργία μίας αρχικής ομαδοποίησης των δεδομένων. Στην πρώτη μέθοδο χρησιμοποιήθηκε ο αλγόριθμος FCM, ενώ στην δεύτερη χρησιμοποιήθηκε το μοντέλο των αυτο-οργανούμενων χαρτών. Στην συνέχεια, και στις δύο περιπτώσεις εφαρμόζονται τεχνικές ιεραρχικής συγχώνευσης, οι οποίες εκμεταλλευόμενες μετα-δεδομένα που προκύπτουν στην κάθε περίπτωση, έχουν ως στόχο την προσέγγιση του βέλτιστου αριθμού των ομάδων.

Με χρήση της δεύτερης μεθοδολογίας αναπτύχθηκε μία μέθοδος αξιολόγησης των χαρακτηριστικών των δεδομένων εισόδου ως προς την συνεισφορά τους στον σχηματισμό των ομάδων. Η μέθοδος αυτή αποτέλεσε και τμήμα ενός πρωτότυπου υβριδικού συστήματος κατηγοριοποίησης που χρησιμοποίησε την μέθοδο αυτή ως στάδιο προεπεξεργασίας και εξαγωγής μετα-δεδομένων, τα οποία αξιοποιήθηκαν από ένα νευρο-ασαφές δίκτυο που αποτελεί το δεύτερο τμήμα του συστήματος. Στην συνέχεια, αναπτύχθηκε ένα ακόμα υβριδικό σύστημα κατηγοριοποίησης του οποίου το κυρίως μέρος αποτελεί ένας πρωτότυπος αλγόριθμος κατηγοριοποίησης που ανήκει στην κατηγορία των αλγόριθμων κατηγοριοποίησης με χρήση ένα σύνολο δεδομένων ως βάση γνώσης. Το δεύτερο τμήμα του συστήματος είναι μία μέθοδος αξιολόγησης των συνδυασμών των χαρακτηριστικών εισόδου με την χρήση αυτο-οργανούμενων χαρτών.

Το μοντέλο των αυτο-οργανούμενων χαρτών χρησιμοποιήθηκε επίσης για την υλοποίηση δύο προσεγγίσεων του μετασχηματισμού της εξαγόμενης από τα δεδομένα γνώσης σε συμβολική μορφή. Η πρώτη προσέγγιση είναι η υλοποίηση

ενός πρωτότυπου υβριδικού συστήματος αναγνώρισης χειρονομιών, στο οποίο οι αυτο-οργανούμενοι χάρτες χρησιμοποιήθηκαν ως παραγωγοί συμβολικών καταστάσεων από τα δεδομένα εισόδου, οι οποίες αποτέλεσαν τα δεδομένα για την δημιουργία πιθανοτικών μοντέλων κατηγοριοποίησης. Στην δεύτερη προσέγγιση παρουσιάστηκαν μεθοδολογίες εξαγωγής συμβολικών κανόνων που αποτελούν και μία μορφή απεικόνισης γνώσης, η οποία είναι κατανοητή και χρηστική από τον άνθρωπο.

Η αποδοτικότητα των μεθόδων που αναπτύχθηκαν αξιολογήθηκε πειραματικά με χρήση συνόλων δεδομένων, τα οποία χρησιμοποιούνται ευρέως από την επιστημονική κοινότητα για την αξιολόγηση μεθόδων και αλγόριθμων των συναφών ερευνητικών πεδίων.

ABSTRACT

The research on hybrid systems of artificial intelligence and machine learning models presents an increasing interest during recent years. The contribution of the current thesis is focused on this field by developing novel models and learning methods and designing original hybrid systems.

More specific, initially, two different clustering methods have been implemented, based on the common idea of using first a simple clustering algorithm for the discovery of an initial group of clusters. In the first method, the FCM algorithm is used while in the second the self-organizing maps model. Following, in both cases, a agglomerative clustering algorithm is applied in order to achieve a more efficient clustering, each time with the necessary modifications so as to take advantage of meta-data deriving from the initial clustering procedure.

With the use of the second clustering procedure, a feature evaluation method has been developed. This method is the first part of a hybrid classification system and the results of the method are fed to a neuro-fuzzy classifier, which is the next part of the system. Another hybrid system has also been developed. The main module of this system is a novel memory-based classifier which interacts with a feature evaluation module that is based on the self-organizing maps model.

The self organizing maps model has been also employed for implementing transformations of extracted knowledge to symbolic form. The first approach of this kind has been the development of a hand gesture recognition system that uses self-organizing maps for generating symbolic states so as to create probabilistic classification models. The second approach is a group of rule extraction methods from a trained self-organizing map.

The performance of all the above methods and systems has been tested on benchmark datasets, which are widely used by researchers in the corresponding fields.

Κεφάλαιο 1

Αντικείμενο της διδακτορικής διατριβής

1.1 Εισαγωγή

Η έρευνα και ανάπτυξη στην περιοχή της εξόρυξης γνώσης από δεδομένα έχει γνωρίσει τεράστια πρόοδο την τελευταία πενταετία, με στόχο την ανάπτυξη μεθόδων και αλγορίθμων που εξάγουν χρήσιμες «κανονικότητες» από μεγάλες ποσότητες δεδομένων, είτε με τη μορφή κανόνων που χαρακτηρίζουν τις σχέσεις μεταξύ μεταβλητών, είτε με τη μορφή συναρτήσεων που επιτρέπουν την ταξινόμηση (classification), παλινδρόμηση (regression) ή αναπαράσταση ιδιοτήτων της κατανομής των δεδομένων. Οι αλγόριθμοι αυτοί πρέπει να χειρίζονται μεγάλα σύνολα δεδομένων υψηλών διαστάσεων χρησιμοποιώντας αποδοτικά την προϋπάρχουσα (a priori) γνώση στην στρατηγική αναζήτησης [11], [22], [32], [34].

Πέραν των παραδοσιακών μεθόδων στατιστικής ανάλυσης, νέες τεχνικές προερχόμενες από τις περιοχές της υπολογιστικής νοημοσύνης (νευρωνικά δίκτυα, ασαφή συστήματα) και της μηχανικής μάθησης συνεισφέρουν με καινοτόμο τρόπο στην ανάλυση των δεδομένων [49], [50]. Η καινοτομία των μεθόδων αυτών μπορεί να συνοψιστεί στο ότι δεν βασίζονται σε στατιστικές υποθέσεις και απλά μαθηματικά μοντέλα, αλλά μπορούν να μάθουν πολύπλοκες μη γραμμικές εξαρτήσεις από τα δεδομένα. Σε αντίθεση με τη συνήθη στατιστική προσέγγιση του ελέγχου θεωριών, τα νευρωνικά δίκτυα και οι συναφείς μέθοδοι στοχεύουν στην επίτευξη ικανοποιητικής γενίκευσης σε νέα δεδομένα ως προς αυτά που χρησιμοποιήθηκαν για την εκπαίδευσή τους. Διάφορες τεχνικές έχουν αναπτυχθεί προς την κατεύθυνση της βελτίωσης της γενικευτικής ικανότητας των νευρωνικών δικτύων. Σε σχέση με τις εφαρμογές εξόρυξης γνώσης, τα

σημαντικότερα προβλήματα κατά την ανάπτυξη ευφυών τεχνικών βασισμένων στα νευρωνικά δίκτυα αφορούν την αντιμετώπιση των υψηλών διαστάσεων των δεδομένων (μεγάλου αριθμού μεταβλητών) και την προσαρμογή σε σύνολα δεδομένων μεγάλης κλίμακας [6]. Επιπλέον, ιδιαίτερη σημασία έχει η εξαγωγή γνώσης σε μορφή εύχρηστη και κατανοητή από τον χρήστη (κανόνες) [5], [74] καθώς και η δυνατότητα οπτικής αναπαράστασης.

Τα τελευταία έτη, επίσης, χαρακτηρίζονται από έντονη ερευνητική δραστηριότητα στον τομέα της ανάπτυξης υβριδικών μοντέλων που βασίζονται στον συνδυασμό υποσυμβολικής (νευρωνικής) και συμβολικής επεξεργασίας (συνήθως βασισμένης σε κανόνες). Η σύνθεση αυτή, η οποία είναι συμβατή με την ετερογενή φύση των ανθρώπινων γνωστικών λειτουργιών, οδηγεί σε μεθόδους που επωφελούνται από την αλληλεπίδραση και τα πλεονεκτήματα των επιμέρους συνιστωσών, δημιουργώντας ισχυρά και ευέλικτα ευφυή συστήματα κατάλληλα για την αναπαράσταση και επεξεργασία τόσο διαδικαστικής όσο και δηλωτικής γνώσης [71], [72].

Η δυνατότητα συνδυασμού μοντέλων χαμηλού και υψηλού επιπέδου είναι ιδιαίτερα σημαντική όσον αφορά την εξόρυξη γνώσης από δεδομένα, καθόσον επιτρέπει την συνέργεια της γενικευτικής ικανότητας των νευρωνικών μοντέλων με τις δυνατότητες επεξήγησης και αιτιολόγησης των αποφάσεων σε κατανοητή μορφή που διαθέτουν τα συστήματα συμβολικής επεξεργασίας [26] , [23].

1.2 Στόχος της διατριβής

Η παρούσα διατριβή έχει ως στόχο να συνεισφέρει στην αντιμετώπιση των βασικών ζητημάτων που αναφέρθηκαν παραπάνω σε σχέση με τις εφαρμογές εξόρυξης γνώσης από δεδομένα, με τη βοήθεια υβριδικών μοντέλων αναπαράστασης και επεξεργασίας της γνώσης. Ειδικότερα, το αντικείμενο της διατριβής αφορά στην ανάπτυξη και διερεύνηση αποδοτικών ευφυών τεχνικών για προβλήματα ταξινόμησης και ομαδοποίησης. Οι τεχνικές αυτές εντάσσονται σε ένα γενικό μοντέλο υβριδικής επεξεργασίας, στο οποίο οι δύο τύποι συνιστωσών θα λειτουργούν ταυτόχρονα και συνεργητικά αλληλεπιδρώντας σε μια συνεχιζόμενη διαδικασία μάθησης - ανταλλάσσοντας πληροφορία μεταξύ τους και με το περιβάλλον. Οι συνιστώσες θα μπορούν να λειτουργούν υπό την επίβλεψη μετα-επεξεργαστή ή να συνεργάζονται με διάφορους τρόπους, π.χ. εκτελώντας διαφορετικές εργασίες ή εκτελώντας την ίδια εργασία με διαφορετικούς τρόπους ή/και υπό διαφορετικές συνθήκες. Το γενικό σχήμα που θα μελετηθεί θα μπορεί να συγκεκριμενοποιηθεί με τη χρήση διάφορων επιμέρους μοντέλων, παρέχοντας εξειδικευμένες τεχνικές με διαφορετικές ιδιότητες κατά περίπτωση.

Τα βασικά θέματα που θα πρέπει να αντιμετωπιστούν κατά την ανάπτυξη και διερεύνηση των υβριδικών ευφυών μοντέλων αφορούν:

- Δομή / αρχιτεκτονική: τμηματικότητα (modularity), βαθμός κατάτμησης (granularity), βαθμός σύζευξης συνιστωσών.
- Αναπαράσταση γνώσης: βαθμός ομοιογένειας, επιλογή χαρακτηριστικών / μείωση διαστάσεων, μετασχηματισμοί, ενσωμάτωση γνώσης, εξαγωγή γνώσης.
- Μάθηση: ενιαία / κατανεμημένη, αλληλεπιδράσεις / συνέργεια μεθόδων εκπαίδευσης, υβριδικοί αλγόριθμοι, υπολογιστική πολυπλοκότητα, προσαρμογή σε μεγάλα σύνολα δεδομένων (scaling), δυνατότητες επανεκπαίδευσης.

Το θέμα της μάθησης αποτελεί την κύρια ερευνητική πρόκληση και αφορά τόσο στην μάθηση του περιεχομένου (γνώσης) όσο και στην μάθηση (δημιουργία) της ίδιας της αρχιτεκτονικής ενός ευφυούς συστήματος. Θα πρέπει να σημειωθεί ότι το πρόβλημα της εκπαίδευσης υβριδικών συστημάτων με ενιαίο τρόπο άρχισε να διερευνάται σχετικά πρόσφατα.

1.3 Συνεισφορά της διατριβής

Η συνεισφορά της διατριβής εντάσσεται αφενός στην περιοχή των υβριδικών συστημάτων με την ανάπτυξη νέων μοντέλων και αλγορίθμων μάθησης και αφετέρου στην περιοχή της εξόρυξης γνώσης από δεδομένα. Η αποδοτικότητα των μεθόδων που αναπτύχθηκαν αξιολογήθηκε πειραματικά με χρήση συνόλων δεδομένων, τα οποία χρησιμοποιούνται ευρέως από την επιστημονική κοινότητα για την αξιολόγηση μεθόδων και αλγόριθμων των συναφών ερευνητικών πεδίων. Ιδιαίτερα σημαντική είναι η συσχέτιση και σύγκριση των μεθόδων αυτών με άλλες τεχνικές ανάλυσης.

Πιο συγκεκριμένα, στο πλαίσιο της διατριβής αναπτύχθηκαν δυο μεθοδολογίες ομαδοποίησης. Στις μεθοδολογίες αυτές κοινό στοιχείο αποτελεί η χρήση αρχικά ενός γνωστού αλγόριθμου ομαδοποίησης για την δημιουργία μίας αρχικής ομαδοποίησης των δεδομένων, η οποία όμως συγκροτείτε από μεγάλο αριθμό ομάδων. Στην πρώτη μέθοδο χρησιμοποιήθηκε ο αλγόριθμος FCM (Fuzzy C-Means), ενώ στην δεύτερη χρησιμοποιήθηκε το μοντέλο των αυτο-οργανούμενων χαρτών (Self-Organizing Maps). Στην συνέχεια, και στις δύο περιπτώσεις, εφαρμόζεται μία τεχνική ιεραρχικής συγχώνευσης των ομάδων αυτών με στόχο την

όσο το δυνατόν καλύτερη προσέγγιση του βέλτιστου αριθμού ομάδων του συνόλου των δεδομένων εισόδου, η οποία όμως είναι κατάλληλα διαμορφωμένη έτσι ώστε να εκμεταλλεύεται μετα-δεδομένα που προκύπτουν στην κάθε περίπτωση. Η πρώτη από τις δύο μεθοδολογίες δημοσιεύτηκε στην εργασία [24] ενώ η δεύτερη στην εργασία [56].

Με χρήση της δεύτερης μεθοδολογίας αναπτύχθηκε μία μέθοδος αξιολόγησης των χαρακτηριστικών των δεδομένων εισόδου ως προς την συνεισφορά τους στον σχηματισμό των ομάδων που σχηματίστηκαν από τα δεδομένα εισόδου. Η μέθοδος αυτή αποτέλεσε και τμήμα ενός πρωτότυπου υβριδικού συστήματος κατηγοριοποίησης. Το σύστημα αυτό χρησιμοποίησε την μέθοδο αυτή ως στάδιο προεπεξεργασίας των δεδομένων και εξαγωγής μετα-δεδομένων τα οποία αξιοποιήθηκαν από ένα νευρο-ασαφές δίκτυο που και αποτελεί το δεύτερο τμήμα του συστήματος. Το υβριδικό αυτό σύστημα παρουσιάστηκε στην εργασία [56].

Στην συνέχεια αναπτύχθηκε ένα ακόμα υβριδικό σύστημα κατηγοριοποίησης του οποίου το κύριο μέρος αποτελεί ένας πρωτότυπος αλγόριθμος κατηγοριοποίησης που ανήκει στην κατηγορία των αλγόριθμων κατηγοριοποίησης με χρήση ένα σύνολο δεδομένων ως βάση γνώσης. Το δεύτερο τμήμα του συστήματος είναι μία μέθοδος αξιολόγησης των συνδυασμών των χαρακτηριστικών εισόδου με την χρήση αυτο-οργανούμενων χαρτών. Από αυτό το σύστημα, κατά τα διάφορα στάδια ανάπτυξής του προέκυψαν οι εργασίες [58],[57],[59].

Το μοντέλο των αυτο-οργανούμενων χαρτών χρησιμοποιήθηκε επίσης για την υλοποίηση δύο προσεγγίσεων του μετασχηματισμού της εξαγόμενης από τα δεδομένα γνώσης σε συμβολική μορφή. Η πρώτη προσέγγιση ήταν η υλοποίηση ενός πρωτότυπου υβριδικού συστήματος αναγνώρισης χειρονομιών. Σε αυτή την προσέγγιση, οι αυτο-οργανούμενοι χάρτες χρησιμοποιήθηκαν ως παραγωγοί συμβολικών καταστάσεων από τα δεδομένα εισόδου, οι οποίες αποτέλεσαν την βάση για την δημιουργία πιθανοτικών μοντέλων κατηγοριοποίησης. Στην δεύτερη προσέγγιση παρουσιάστηκαν μεθοδολογίες εξαγωγής συμβολικών κανόνων που αποτελούν και μία μορφή απεικόνισης της εξαγόμενης από τα δεδομένα γνώσης, η οποία είναι κατανοητή και χρηστική από τον άνθρωπο. Οι δύο αυτές προσεγγίσεις παρουσιάζονται στις εργασίες [10] και [55] αντίστοιχα.

1.4 Διάρθρωση της διατριβής

Στο επόμενο κεφάλαιο γίνεται μία σύντομη επισκόπηση των υβριδικών συστημάτων, καθώς και του μοντέλου των αυτο-οργανούμενων χαρτών και της μεθοδολογίας κατηγοριοποίησης των «κ πλησιέστερων γειτόνων». Στο κεφάλαιο 3 παρουσιάζονται οι τεχνικές που έχουν αναπτυχθεί για την ομαδοποίηση δε-

δομένων. Στο κεφάλαιο 4 γίνεται περιγραφή του πρωτότυπου υβριδικού νευρο-ασαφούς κατηγοριοποιητή και στο κεφάλαιο 5 παρουσιάζεται το σύστημα κατηγοριοποίησης βασισμένο στο μοντέλο των κ-πλησιέστερων γειτόνων. Στο κεφάλαιο 6 παρουσιάζονται οι δύο προσεγγίσεις για την παραγωγή γνώσης σε συμβολικής μορφή. Τέλος, στο κεφάλαιο 7 γίνεται μια συνολική αναφορά στα συμπεράσματα που προέκυψαν από την εκπόνηση της διατριβής καθώς και σε θέματα για μελλοντική εργασία.

□

Κεφάλαιο 2

Εισαγωγικά θέματα

Στο κεφάλαιο αυτό περιγράφονται συνοπτικά θέματα τα οποία κρίνονται θεμελιώδη για την έρευνα που πραγματοποιήθηκε στο πλαίσιο της διατριβής αυτής. Στην πρώτη ενότητα παρουσιάζεται μία συνοπτική περιγραφή των κατηγοριών των υβριδικών συστημάτων. Στην δεύτερη ενότητα περιγράφεται το μοντέλο των αυτο-οργανούμενων χαρτών, ενώ στην τρίτη ο αλγόριθμος των κ-πλησιέστερων γειτόνων. Τέλος, γίνεται μία σύντομη αναφορά σε μεθόδους αξιολόγησης των χαρακτηριστικών-εισόδων ενός συνόλου δεδομένων.

2.1 Επισκόπηση υβριδικών συστημάτων

Η ανάπτυξη υβριδικών συστημάτων στον χώρο της τεχνητής νοημοσύνης και της μηχανικής μάθησης βρίσκεται τα τελευταία χρόνια ανάμεσα στα πιο δημοφιλή πεδία έρευνας. Η ανάπτυξη ενός τέτοιου συστήματος πραγματοποιείται με την ολοκλήρωση διαφορετικών τεχνικών μάθησης και προσαρμογής με στόχο να αναπτυχθούν αλληλεπιδράσεις μέσω του υβριδισμού και της συγχώνευσης, οι οποίες θα βοηθήσουν να υπερκεραστούν οι περιορισμοί των επιμέρους τεχνικών.

Οι κατηγορίες συστημάτων που μετέχουν στη δημιουργία των υβριδικών συστημάτων είναι οι εξής: τα τεχνητά νευρωνικά δίκτυα, τα συστήματα ασαφούς λογικής και οι αλγόριθμοι βελτιστοποίησης (γενετικοί αλγόριθμοι, εξελικτικός προγραμματισμός). Στον πίνακα 2.1 φαίνονται τα κύρια πλεονεκτήματα των συστημάτων αυτών.

Στον πίνακα 2.2 περιγράφονται τα βασικά χαρακτηριστικά των συστημάτων συμβολικής και υποσυμβολικής επεξεργασίας τα οποία αποτελούν συνήθως τα δομικά στοιχεία ενός υβριδικού μοντέλου.

Τα υβριδικά συστήματα μπορούν να κατηγοριοποιηθούν με βάση τον βαθμό αλληλεπίδρασης των μερών που τα συνιστούν [2], [48], [82].

Πίνακας 2.1: *Πλεονεκτήματα συστημάτων*

Σύστημα	Πλεονεκτήματα
Νευρωνικά δίκτυα	Προσαρμοστικότητα, Μάθηση, Γενίκευση
Ασαφής λογική	Προσεγγιστική Λογική
Αλγόριθμοι βελτιστοποίησης	Βελτιστοποίηση χωρίς χρήση παραγώγων

Πίνακας 2.2: *Χαρακτηριστικά συστημάτων υποσυμβολικής και συμβολικής επεξεργασίας*

	Συστήματα υποσυμβολικής επεξεργασίας	Συστήματα συμβολικής επεξεργασίας
Αναπαράσταση γνώσης	Συνδέσεις-δομή δικτύου	Κανόνες, Καταστάσεις
Στοιχεία υπολογισμού	Κόμβοι, Βάρη	Λογικές προτάσεις, Πιθανότητες
Επεξεργασία	Συνεχείς τιμές	Διακριτά Σύμβολα

Συστήματα από μη αλληλεπιδρώντα μοντέλα

Στην κατηγορία αυτή εντάσσονται συστήματα τα οποία αποτελούνται από αυτόνομα μοντέλα χωρίς καμία αλληλεπίδραση μεταξύ τους. Ο σκοπός δημιουργίας τέτοιων συστημάτων είναι η σύνθεση της λύσης από λύσεις που δίνουν τα επιμέρους μοντέλα σε υποπροβλήματα του αρχικού προβλήματος ή σύγκριση των αποτελεσμάτων των επιμέρους μοντέλων σε ένα κοινό περιβάλλον λειτουργίας. Τέτοια υβριδικά συστήματα παρουσιάζουν τον μικρότερο βαθμό υβριδικότητας.

Συστήματα μετασχηματισμού γνώσης

Τα συστήματα αυτής της κατηγορίας χρησιμοποιούν μοντέλα υποσυμβολικής ή συμβολικής επεξεργασίας για να μετασχηματίσουν την γνώση που εξάγεται από τα δεδομένα από την μία στην άλλη μορφή. Αυτό συμβαίνει όταν το πρόβλημα είναι ιδανικό για επίλυση από μία κατηγορία μοντέλων αλλά τα αποτελέσματα ζητούνται σε μία μορφή που μπορεί να δώσει μία άλλη κατηγορία μοντέλων ή τα δεδομένα είναι σε μη κατάλληλη μορφή.

Συστήματα με ιεραρχική αρχιτεκτονική

Σε αυτήν την κατηγορία ανήκουν υβριδικά συστήματα τα οποία χαρακτηρίζονται από μία ιεραρχική δόμηση των επιμέρους μοντέλων που τα συγκροτούν. Δηλαδή τα προβλήματα που καλούνται να αντιμετωπίσουν τα συστήματα αυτά, υφίστα-

νται διαδοχική επεξεργασία από τα διαφορετικά μοντέλα του συστήματος και τα αποτελέσματα της επεξεργασίας του ενός μοντέλου τροφοδοτούν το μοντέλο που βρίσκεται στην επόμενη βαθμίδα επεξεργασίας.

Συστήματα από αλληλεπιδρώντα μοντέλα

Η κατηγορία αυτή περιλαμβάνει συστήματα όπου τα επιμέρους μοντέλα ανταλλάσσουν πληροφορίες αμφίδρομα κατά την επεξεργασία των δεδομένων και την διαδικασία επίλυσης του προβλήματος. Η ροή των δεδομένων εισόδου μπορεί να είναι είτε παράλληλη σε περισσότερα από ένα επιμέρους μοντέλα είτε ιεραρχική όπως στην προηγούμενη κατηγορία αλλά με την διαφορά ότι μπορεί υπάρχει και ανατροφοδότηση πληροφοριών από μοντέλο υψηλότερης ιεραρχικά βαθμίδας προς ένα άλλο χαμηλότερης βαθμίδας.

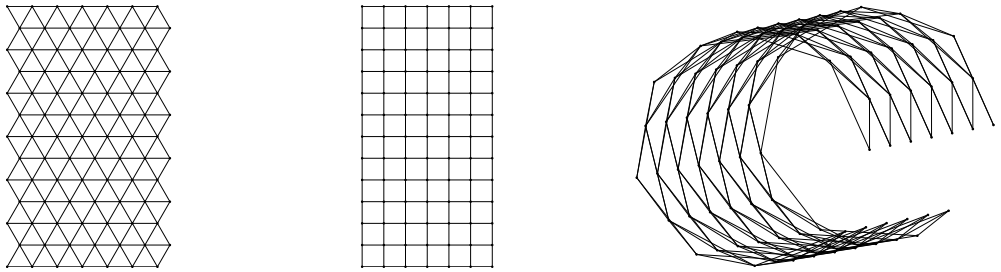
Ολοκληρωμένα υβριδικά συστήματα

Αυτή η κατηγορία αποτελείται από υβριδικά συστήματα στα οποία τα συμμετέχοντα μοντέλα έχουν συντηχθεί και έχουν συντελέσει στην δημιουργία ενός καινούργιου υπολογιστικού μοντέλου. Αυτό συμβαίνει όταν τα επιμέρους μοντέλα έχουν αποσυντεθεί και τα μέρη των μοντέλων αυτών δημιουργούν το νέο υβριδικό σύστημα. Στην αρχιτεκτονική του προκύπτοντος συστήματος δεν αναγνωρίζονται τα συνιστώντα μοντέλα σαν αυτοτελή μέρη αλλά μόνο κομμάτια ή μεθοδολογίες αυτών. Τα συστήματα της κατηγορίας αυτής παρουσιάζουν τον μεγαλύτερο βαθμό υβριδικότητας.

2.2 Αυτο-οργανούμενοι Χάρτες (Self Organizing Maps - SOM)

Ένας αυτο-οργανούμενος χάρτης (SOM)[41],[42] αποτελείται από ένα πλέγμα νευρώνων συνήθως μίας ή δύο διαστάσεων. Πλέγματα μεγαλύτερων διαστάσεων μπορούν επίσης να χρησιμοποιηθούν αλλά δεν αποτελεί συνήθη πρακτική. Η συνδεσμολογία των νευρώνων στη περίπτωση που το πλέγμα είναι δύο διαστάσεων ποικίλει, στο σχήμα 2.1 (α) και (β) παρουσιάζονται οι δύο περισσότερο χρησιμοποιούμενες συνδεσμολογίες, η εξαγωνική και η ορθογώνια. Τα πλέγματα εκτός από την συνδεσμολογία μπορούν να ποικίλουν και ως προς την μορφή (σχήμα 2.1 (γ)).

Οι νευρώνες του πλέγματος αποτελούνται από διανύσματα διάστασης ίσης με την διάσταση του χώρου των δεδομένων εκπαίδευσης. Η εκπαίδευση πραγματο-



(α) (β) (γ)

Σχήμα 2.1: *Είδη πλέγματος αυτο-οργανούμενων χαρτών. (α) Εξαγωνικό επίπεδο πλέγμα. (β) Ορθογώνιο επίπεδο πλέγμα. (γ) Κυλινδρικό πλέγμα.*

ποιείται σε δύο φάσεις, την φάση του ανταγωνισμού και την φάση της συνεργασίας. Στην φάση του ανταγωνισμού, για κάθε πρότυπο εισόδου, ο αλγόριθμος υπολογίζει την απόσταση του προτύπου από τα διανύσματα που αντιστοιχούν στους νευρώνες. Για τον υπολογισμό της απόστασης χρησιμοποιείται η μετρική της ευκλείδειας απόστασης χωρίς να αυτό αποκλείει και την χρήση άλλων μετρικών. Στην φάση αυτή αναδεικνύεται νικητής ο νευρώνας με την μικρότερη απόσταση από το πρότυπο εισόδου.

Στη συνέχεια εκτελείται η φάση την συνεργασίας κατά την οποία γίνεται η ενημέρωση των διανυσμάτων των νευρώνων. Οι νευρώνες των οποίων τα διανύσματα θα ενημερωθούν ορίζονται με βάση την θέση του νευρώνα νικητή στο πλέγμα του χάρτη και με την χρήση μίας συνάρτησης γειτνίασης που ορίζει έναν αριθμό νευρώνων γύρω από τον νευρώνα νικητή των οποίων τα διανύσματα θα ενημερωθούν επίσης.

2.2.1 Φάση ανταγωνισμού

Έστω ότι το m είναι ίσο με την διάσταση του χώρου των δεδομένων εισόδου και ένα από τα πρότυπα εισόδου είναι το $\mathbf{x}$ με $\mathbf{x} = [x_1, x_2, ..., x_m]$. Το διάνυσμα του νευρώνα j του πλέγματος ορίζεται ως :

$$\mathbf{w}_j = [w_{j1}, w_{j2}, ..., w_{jm}], \quad j = 1, 2, ..., k$$

όπου k ο συνολικός αριθμός των νευρώνων του πλέγματος. Στην φάση αυτή εντοπίζεται ο νευρώνας νικητής. Ως $i(\mathbf{x})$ ορίζεται η συνάρτηση που αναδεικνύει τον νευρώνα νικητή για το πρότυπο εισόδου $\mathbf{x}$.

$$i(\mathbf{x}) = \arg\min_j(dist(\mathbf{x}, \mathbf{w}_j)), \quad j = 1, 2, ..., k \tag{2.1}$$

Η συνάρτηση dist υποδηλώνει την μετρική που χρησιμοποιείται για τον υπολογισμό της απόστασης μεταξύ των διανυσμάτων. Ο νευρώνας νικητής καλείται και νευρώνας μεγαλύτερης ομοιότητας (Best Matching Unit - BMU). Η

διαδικασία αυτή μπορεί να περιγραφεί ως απεικόνιση του συνεχούς χώρου των δεδομένων εισόδου στον διακριτό χώρο $1, 2, k$ που αντιστοιχεί στους νευρώνες του πλέγματος.

2.2.2 Φάση συνεργασίας

Ο νικητής νευρώνας ορίζει επάνω στο πλέγμα το κέντρο μίας τοπολογικής περιοχής - γειτονίας, η οποία περιλαμβάνει τους συνεργαζόμενους νευρώνες. Η γειτονία αυτή ορίζεται μέσω μίας συνάρτησης γειτνίασης του μεταξύ του νευρώνα νικητή και των υπολοίπων νευρώνων επάνω στο πλέγμα. Η συνάρτηση αυτή πρέπει να πληρεί τις εξής προϋποθέσεις.

- Να είναι συμμετρική γύρω από την θέση του νευρώνα νικητή.
- Να λαμβάνει την μέγιστη τιμή της στο κέντρο της γειτονίας, δηλαδή στην θέση του νευρώνα νικητή.
- Οι τιμές τις πρέπει να φθίνουν μονότονα καθώς αυξάνεται η απόσταση των νευρώνων από τον νευρώνα νικητή.

Η απόσταση μεταξύ δύο νευρώνων στο πλέγμα δεν ορίζεται από την απόσταση των διανυσμάτων τους αλλά είναι συνάρτηση της θέσης τους επάνω στο πλέγμα. Έστω $h_{j,i}$ η συνάρτηση που ορίζει την γειτονία γύρω από τον νευρώνα νικητή i και j ένας οποιοσδήποτε νευρώνας του πλέγματος. Αν ορίσουμε ως $d_{j,i}$ την απόσταση μεταξύ των νευρώνων i και j τότε μία τυπική επιλογή που πληρεί και τις παραπάνω προϋποθέσεις είναι η συνάρτηση Gauss.

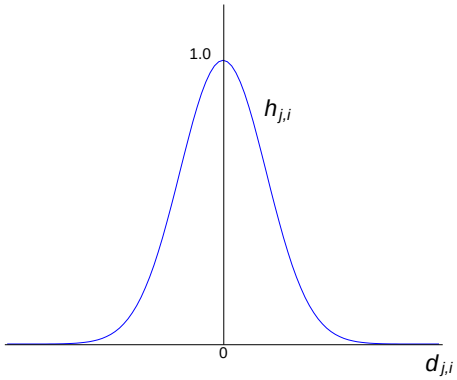
$$h_{j,i(\mathbf{x})} = \exp(-\frac{d_{j,i}^2}{2\sigma^2}) \tag{2.2}$$

Η παραπάνω συνάρτηση είναι ανεξάρτητη από την θέση του νευρώνα νικητή. Η παράμετρος σ ορίζει την ακτίνα επιρροής της συνάρτησης άρα και το μέγεθος της γειτονίας.

Η απόσταση $d_{j,i}$ ορίζεται ως :

$$d_{j,i}^2 = \|r_i - r_j\|^2 \tag{2.3}$$

όπου r_i και r_j είναι τα διακριτά διανύσματα που ορίζουν την θέση επάνω στο πλέγμα του νευρώνα νικητή i και ενός νευρώνα j αντίστοιχα στην περίπτωση που το πλέγμα είναι επίπεδο, δύο διαστάσεων. Στην πιο απλή περίπτωση ενός μονοδιάστατου πλέγματος η απόσταση μπορεί να οριστεί ως : $d_{j,i} = \|r_i - r_j\|$.



Σχήμα 2.2: *Γραφική παράσταση της συνάρτησης Gauss*

Κατά τη διάρκεια της εκπαίδευσης του χάρτη, οι δύο φάσεις επαναλαμβάνονται κάθε φορά που στον χάρτη παρουσιάζεται ένα νέο δεδομένο εισόδου. Όταν όλα τα δεδομένα παρουσιαστούν στον χάρτη τότε η διαδικασία επαναλαμβάνεται για όλο το σύνολο των δεδομένων εισόδου.

Κατά την διάρκεια της εκπαίδευσης, το μέγεθος της γειτονίας γύρω από τον νευρώνα νικητή μειώνεται. Στην περίπτωση που η συνάρτηση γειτονιάς είναι η συνάρτηση Gauss, τότε η μείωση αυτή επιτυγχάνεται με μείωση της παραμέτρου σ. Επομένως η παράμετρος αυτή γίνεται συνάρτηση του χρόνου. Η παρακάτω εκθετική συνάρτηση είναι μία τυπική επιλογή για τον έλεγχο της παραμέτρου σ

$$\sigma(n) = \sigma_0 \exp(-\frac{n}{\tau}), \quad n = 0, 1, 2, ... \tag{2.4}$$

όπου n είναι η διακριτή μεταβλητή του χρόνου και είναι ίση με τον τρέχοντα αριθμό των εποχών κατά την διάρκεια της εκπαίδευσης, ενώ η παράμετρος τ είναι μία σταθερά. Με την χρήση της παραπάνω συνάρτησης, η συνάρτηση γειτνίασης μετατρέπεται και σε συνάρτηση του χρόνου.

$$h_{j,i(\mathbf{x})}(n) = \exp(-\frac{d_{j,i}^2}{2\sigma^2(n)}), \quad n = 0, 1, 2, ... \tag{2.5}$$

Κατά τη φάση της συνεργασίας τα διανύσματα των νευρώνων που ανήκουν στην γειτονία του νευρώνα νικητή πρέπει να τροποποιηθούν έτσι ώστε να πραγματοποιηθεί ή διαδικασία της αυτο-οργάνωσης του χάρτη. Η τροποποίηση αυτή έχει ως στόχο την μείωση της απόστασης των νευρώνων της γειτονίας από το εκάστοτε πρότυπο εισόδου. Επομένως απαιτείται η κατάλληλη τροποποίηση των διανυσμάτων των νευρώνων. Η τροποποίηση αυτή προκύπτει από την παρακάτω εξίσωση

$$\Delta \mathbf{w}_j = a h_{j,i(\mathbf{x})}(\mathbf{x} - \mathbf{w}_j) \tag{2.6}$$

όπου a είναι η ρυθμός μάθησης και χρησιμοποιείται για να ελέγξει το ποσοστό της μεταβολής του διανύσματος του εκάστοτε νευρώνα. Και η παράμετρος αυτή πρέπει να είναι συνάρτηση του χρόνου έτσι ώστε η διαδικασία αυτό-διοργάνωσης του χάρτη να συγκλίνει σε μία σταθερή λύση. Επομένως ο ρυθμός μάθησης πρέπει να είναι αυξημένος στην αρχή της διαδικασίας έτσι ώστε να επιτρέπει μεγάλες διαταραχές στα διανύσματα των νευρώνων και στην συνεχεία να φθίνει μονότονα έτσι ώστε οι αλλαγές αυτές να περιοριστούν και να οδηγήσουν σε μία σταθερή δομή του χάρτη. Εισάγοντας και αυτή την παράμετρο, προκύπτει ότι το νέο διάνυσμα του νευρώνα j θα είναι :

$$\mathbf{w}_j(n+1) = \mathbf{w}_j(n) + a(n)h_{j,i(\mathbf{x})}(n)(\mathbf{x} - \mathbf{w}_j(n)) \tag{2.7}$$

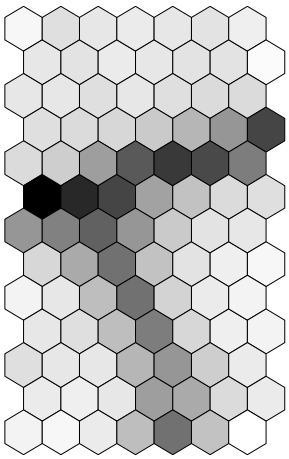
2.2.3 Ιδιότητες αυτο-οργανούμενων χαρτών

Ένας χάρτης SOM, μετά την εκπαίδευσή του, συγκεντρώνει την σημαντικότερη πληροφόρηση από τα δεδομένα εισόδου. Δεδομένα, τα οποία παρουσιάζουν ομοιότητα ως προς τα χαρακτηριστικά τους, συγκεντρώνονται στον ίδιο νευρώνα του χάρτη. Γειτονικοί νευρώνες στο πλέγμα του χάρτη έχουν την ιδιότητα να συγκεντρώνουν δεδομένα τα οποία επίσης παρουσιάζουν ομοιότητα, αλλά σε βαθμό μικρότερο από ότι εκείνη που παρουσιάζεται μεταξύ των προτύπων του ίδιου νευρώνα. Η ιδιότητα αυτή είναι μεταβλητή κατά την έκταση του πλέγματος του χάρτη, δηλαδή η ομοιότητα μεταξύ δεδομένων γειτονικών νευρώνων δεν παρουσιάζεται σε όλες τις περιοχές του χάρτη στον ίδιο βαθμό. Το στοιχείο αυτό επιτρέπει να προσδιοριστούν περιοχές στον χάρτη που συγκροτούν ευρύτερες ομάδες όμοιων δεδομένων. Το πρόβλημα έγκειται στον αυτόματο εντοπισμό των περιοχών αυτών.

Στο σχήμα 2.3 φαίνεται το παράδειγμα ενός αυτο-οργανούμενου χάρτη μετά την εκπαίδευση. Ο μέσος όρος της απόστασης του κάθε νευρώνα από τους νευρώνες με τους οποίους συνδέεται στο πλέγμα αντικατοπτρίζεται από το χρώμα του. Το άσπρο αντιστοιχεί στην ελάχιστη τιμή μεταξύ όλων των μέσων όρων και το μαύρο στην μέγιστη. Οι αποστάσεις υπολογίζονται μεταξύ των διανυσμάτων των νευρώνων στον χώρο εισόδου. Όπως φαίνεται και στο σχήμα υπάρχουν τρεις περιοχές στο πλέγμα όπου φαίνεται ότι οι νευρώνες εμφανίζονται να έχουν μικρές σχετικά αποστάσεις μεταξύ τους. Οι περιοχές αυτές οριοθετούνται από κάποιους νευρώνες οι οποίοι χρωματίζονται με πιο σκούρα χρώματα.

Στην βιβλιογραφία έχουν προταθεί αρκετές μεθοδολογίες για την λύση του παραπάνω προβλήματος, όπως για παράδειγμα η εκτέλεση αλγορίθμων ομαδοποίησης δεδομένων, θεωρώντας όμως δεδομένα εισόδου τα χαρακτηριστικά των

νευρώνων του εκπαιδευμένου χάρτη [35], [80]. Ένα άλλο ζητούμενο στη όλη επεξεργασία των δεδομένων είναι η δυνατότητα αξιολόγησης των χαρακτηριστικών των δεδομένων εισόδου από έναν εκπαιδευμένο χάρτη SOM [63], [78], [37].



Σχήμα 2.3: *Παράδειγμα εκπαιδευμένου χάρτη στο οποίο αποτυπώνεται χρωματικά η απόσταση μεταξύ των γειτονικών νευρώνων.*

2.3 Αλγόριθμος κ-πλησιέστερων γειτόνων

Η βασική ιδέα του αλγόριθμου των «κ - πλησιέστερων γειτόνων» είναι η εύρεση ενός συνόλου προτύπων από μία βάση γνώσης, τα οποία - με βάση τη σύγκριση των χαρακτηριστικών - είναι πλησιέστερα σε ένα πρότυπο άγνωστης κατηγορίας. Η κατηγορία του άγνωστου προτύπου προκύπτει από την πλειοψηφούσα κατηγορία στο σύνολο των γειτονικών προτύπων με βάση τις κατηγορίες των προτύπων αυτών. Η εύρεση των πλησιέστερων προτύπων γίνεται υπολογίζοντας την απόσταση με χρήση μιας μετρικής, συνήθως της ευκλείδειας, [3], [13], [15], [69]. Πολλές παραλλαγές του μοντέλου έχουν προταθεί χρησιμοποιώντας διαφορετικές μετρικές υπολογισμού τις απόστασης [87].

Έστω ότι $\mathbf{x}$ είναι ένα πρότυπο άγνωστης κατηγορίας και D είναι το σύνολο των προτύπων γνωστής κατηγορίας, τα οποία αποτελούν την βάση γνώση. Ο αλγόριθμος σχηματίζει το σύνολο K^x των k πλησιέστερων γειτόνων, υποσύνολο του συνόλου D. Η εύρεση αυτών των προτύπων γίνεται με χρήση της παρακάτω μετρικής.

$$Dist\left(\mathbf{x},\mathbf{y}\right)=\sqrt{\sum_{i}w_i\cdot diff\left(x_i,y_i\right)^2} \tag{2.8}$$

Όπου x_i και y_i είναι οι τιμές του i-οστού χαρακτηριστικού των προτύπων x και y αντίστοιχα και w_i είναι μία παράμετρος στάθμισης του χαρακτηριστικού i. Η συνάρτηση $diff$ που καθορίζει την διαφορά μεταξύ τις τιμές των χαρακτηριστικών μπορεί να οριστεί στη πιο απλή μορφή της ως:

$$diff\left(x_i, y_i\right) = \begin{cases} |x_i - y_i| & \text{, if feature } i \text{ is numeric} \\ 0 & \text{, if feature } i \text{ is symbolic and } x_i = y_i \\ 1 & \text{, otherwise} \end{cases} \tag{2.9}$$

Μια γενική μορφή του κανόνα κατηγοριοποίησης για την εκτίμηση της κατηγορίας $C_{\mathbf{x}}$ του προτύπου $\mathbf{x}$ του αλγόριθμου μπορεί να περιγραφεί ως:

$$C_{\mathbf{x}} = \arg\max_{j} \sum_{\forall \mathbf{y}: C_{\mathbf{y}} = j} g_{\mathbf{x}}\left(\mathbf{y}\right) \tag{2.10}$$

όπου ο δείκτης j αντιστοιχεί στις κατηγορίες όλων των προτύπων του συνόλου D.

Η απλή μορφή του κανόνα κατηγοριοποίησης μπορεί να διατυπωθεί χρησιμοποιώντας ως συνάρτηση g την παρακάτω:

$$g_{\mathbf{x}}\left(\mathbf{y}\right) = \begin{cases} 1 & \text{, if } \mathbf{y} \in K^{\mathbf{x}} \\ 0 & \text{, if } \mathbf{y} \notin K^{\mathbf{x}} \end{cases} \tag{2.11}$$

όπου $K^{\mathbf{x}}$ είναι το σύνολο των k πλησιέστερων γειτόνων στον πρότυπο $\mathbf{x}$. Σε αυτή την απλή μορφή κάθε γείτονας συνεισφέρει ισοδύναμα στον προσδιορισμό την κατηγορίας του άγνωστου πρότυπου. Μια άλλη πολύ διαδεδομένη μορφή της συνάρτησης g είναι η εξής:

$$g_{\mathbf{x}}\left(\mathbf{y}\right) = \begin{cases} \dfrac{1}{Dist\left(\mathbf{x}, \mathbf{y}\right)} & \text{, if } \mathbf{y} \in K^{\mathbf{x}} \\ 0 & \text{, if } \mathbf{y} \notin K^{\mathbf{x}} \end{cases} \tag{2.12}$$

Η παραπάνω μορφή της εξίσωσης g σταθμίζει την συνεισφορά των γειτονικών προτύπων στην τελική απόφαση της κατηγοριοποίησης με βάση την απόστασή τους από το άγνωστο πρότυπο.

2.3.1 Επιλογή-Αξιολόγηση χαρακτηριστικών εισόδου

Παρόλο που οι αλγόριθμοι «μνήμης» συχνά αποκαλούνται και «τεμπέλικοι μαθητές» (lazy learning) λόγω της έλλειψης μίας διαδικασίας προεπεξεργασίας και μάθησης των δεδομένων, ένα σημαντικό μέρος της βιβλιογραφίας είναι αφιερωμένο

στην περιγραφή μεθόδων προεπεξεργασίας. Οι μέθοδοι αυτές χρησιμοποιούνται για την δημιουργία σχημάτων επιλογής ή ισοστάθμισης των χαρακτηριστικών εισόδου κατά τους μετέπειτα υπολογισμούς των αποστάσεων από κάποιο άγνωστο πρότυπο.

Το πρόβλημα βέβαια της επιλογής χαρακτηριστικών είναι γενικότερο και ανεξάρτητο από το πεδίο των αλγορίθμων «μνήμης». Μία μεθοδολογία αξιολόγησης χαρακτηριστικών εάν εξεταστεί υπό το πρίσμα της αλληλεπίδρασής της με τον αλγόριθμο κατηγοριοποίησης που θα χρησιμοποιηθεί σε δεύτερο στάδιο, μπορεί να ενταχθεί σε μία από τις τρεις παρακάτω κατηγορίες.

Πρώτη είναι η κατηγορία των μεθόδων φίλτρων (filter). Στην κατηγορία αυτή ανήκουν οι μεθοδολογίες, οι οποίες εκτελούνται αυτόνομα και ανεξάρτητα από το επόμενο στάδιο επεξεργασίας. Δεν λαμβάνουν υπ'όψιν τους τα αποτελέσματα εκτέλεσης του αλγόριθμου κατηγοριοποίησης που έπεται.

Στην δεύτερη κατηγορία ανήκουν οι μεθοδολογίες περιτύλιξης (wrappers). Οι μεθοδολογίες αυτές επεξεργάζονται τα δεδομένα, παράγουν σχήματα αξιολόγησης τα οποία στην συνέχεια χρησιμοποιούνται από κάποιον αλγόριθμο κατηγοριοποίησης. Τα αποτελέσματα της κατηγοριοποίησης επανατροφοδοτούνται στην μεθοδολογία αξιολόγησης έτσι ώστε να προκύψουν νέα, βελτιωμένα σχήματα αξιολόγησης. Η επανάληψη αυτή σε ορισμένες περιπτώσεις υλοποιείται και σαν σχήμα μάθησης του τελικού σχήματος αξιολόγησης [76]. Οι μεθοδολογίες αυτές παρόλο που αξιοποιούν τα αποτελέσματα της κατηγοριοποίησης, αντιμετωπίζουν τον αλγόριθμο κατηγοριοποίησης ως «μαύρο κουτί». Δηλαδή δεν λαμβάνουν υπ'όψιν τους το είδος και την φύση αυτού του αλγόριθμου, παρά μόνο τα αποτελέσματά του.

Η τρίτη κατηγορία είναι οι ενσωματωμένες (embedded) μεθοδολογίες. Οι μεθοδολογίες αυτές σχεδιάζονται για έναν συγκεκριμένο αλγόριθμο ή μία ομάδα συγγενών αλγορίθμων. Πολλές φορές είναι τμήμα του ίδιου του αλγόριθμου. Οι μεθοδολογίες αυτής της κατηγορίας εκμεταλλεύονται συγκεκριμένα χαρακτηριστικά του αλγόριθμου ταξινόμησης και όχι απλώς τα αποτελέσματα της κατηγοριοποίησης. Μια συστηματική και λεπτομερής καταγραφή μεθοδολογιών που ανήκουν σε όλες τις παραπάνω κατηγορίες έχει γίνει στην αναφορά [30].

Μια σημαντική πλευρά των μεθοδολογιών αξιολόγησης των χαρακτηριστικών των δεδομένων εισόδου είναι ο τρόπος αντιμετώπισης των συσχετίσεων που μπορεί να υπάρχουν μεταξύ των χαρακτηριστικών αυτών. Ο διαχωρισμός μπορεί να γίνει με βάση το εάν η μέθοδος επεξεργάζεται κάθε ένα χαρακτηριστικό ανεξάρτητα από το άλλο (univariate) ή προσπαθεί να αξιολογήσει τα χαρακτηριστικά συνδυαστικά μεταξύ τους (multivariate).

Οι μεθοδολογίες που χρησιμοποιούν την πρώτη προσέγγιση, που δηλαδή υιο-

θετούν την υπόθεση της ανεξαρτησίας μεταξύ των χαρακτηριστικών ως προς την επίλυση του προβλήματος, θεωρούνται αποτελεσματικές και εύρωστες, ιδίως σε περιπτώσεις δεδομένων υψηλών διαστάσεων. Δυστυχώς όμως αυτές οι μεθοδολογίες αδυνατούν να αξιολογήσουν σωστά χαρακτηριστικά τα οποία όταν εξεταστούν αυτόνομα κρίνονται ως μη σχετικά με το εκάστοτε πρόβλημα, όταν όμως αξιολογηθούν σε συνδυασμό με άλλα χαρακτηριστικά τότε προκύπτει ότι ο συνδυασμός τους έχει ιδιαίτερη βαρύτητα στην επίλυση του προβλήματος. Ένα άλλο σημείο στο οποίο αυτές οι μεθοδολογίες αδυνατούν να συνεισφέρουν είναι στον εντοπισμό περιπτώσεων πλεοναζόντων χαρακτηριστικών, δηλαδή χαρακτηριστικών που μεταφέρουν την ίδια ακριβώς πληροφορία και η ύπαρξη τους αποτελεί πλεονασμό και μπορούν να αντικατασταθούν από ένα χαρακτηριστικό που θα μεταφέρει αυτήν την πληροφορία.

Η κατηγορία των μεθόδων που προσπαθεί να αξιολογήσει συνδυασμούς χαρακτηριστικών και όχι μεμονωμένα χαρακτηριστικά, συνήθως εμφανίζουν καλύτερα αποτελέσματα. Όμως στις περισσότερες περιπτώσεις χαρακτηρίζονται από υψηλό υπολογιστικό κόστος, γεγονός που τις καθιστά πρακτικά ανεφάρμοστες σε προβλήματα με δεδομένα υψηλών διαστάσεων. Αυτές οι μεθοδολογίες αποκαλούνται επίσης και μεθοδολογίες επιλογής υποομάδων χαρακτηριστικών. Σαν τέτοιες μεθοδολογίες όμως μπορούν να χαρακτηριστούν και οι μεθοδολογίες της πρώτης κατηγορίας εφόσον παρέχουν μία ταξινομημένη αξιολόγηση των χαρακτηριστικών, οπότε η επιλογή της υποομάδας μπορεί να γίνει με επιλογή των χαρακτηριστικών με την υψηλότερη αξιολόγηση. Η διαφορά βέβαια είναι ότι στην πρώτη περίπτωση η επιλογή της υποομάδας γίνεται γιατί υπάρχει ή αξιολόγηση ότι η συγκεκριμένη σύνθεση των χαρακτηριστικών δίνει καλύτερα αποτελέσματα ενώ στην δεύτερη περίπτωση απλώς επιλέγονται τα καλύτερα μεμονωμένα χαρακτηριστικά.

Μεθοδολογίες ανεξαρτήτων μεταβλητών τύπου φίλτρου χρησιμοποιούνται κυρίως στην στατιστική ανάλυση όπως η Pearson correlation coefficient. Η οικογένεια των αλγορίθμων Relief [64] αποτελεί ένα χαρακτηριστικό παράδειγμα πολυμεταβλητών μεθόδων τύπου φίλτρου. Ενώ αντίθετα μεθοδολογίες περιτύλιξης έχουν υλοποιηθεί για παράδειγμα με την χρήση γενετικών αλγορίθμων [66],[90].

Στο πεδίο των αλγορίθμων μνήμης, οι μεθοδολογίες αξιολόγησης των χαρακτηριστικών εισόδου βελτιώνουν την απόδοση τους [12],[29],[83]. Στην εργασία [83], οι συγγραφείς προτείνουν ένα πολυδιάστατο πλαίσιο μέσω του οποίου μπορούν να κατηγοριοποιηθούν μεθοδολογίες αξιολόγησης. Μια από τις διαστάσεις του πλαισίου κατηγοριοποιεί τις μεθοδολογίες με βάση την γενικότητα του σχήματος αξιολόγησης. Το παραγόμενο σχήμα αξιολόγησης μπορεί να είναι ένα και

γενικό παρέχοντας έναν βαθμό αξιολόγησης για κάθε χαρακτηριστικό, μπορεί όμως να επεκταθεί έως το επίπεδο της εξαγωγής διαφορετικού βαθμού αξιολόγησης για διαφορετικές περιοχές του πεδίου τιμών του κάθε χαρακτηριστικού ή έως και διαφορετικό βαθμό αξιολόγησης για κάθε πρότυπο αποθηκευμένο στην μνήμη.

Επίσης τα πρότυπα που υπάρχουν στην βάση γνώσης, μπορούν να αξιολογηθούν και για κάποια από αυτά είναι δυνατόν να μειωθεί ο βαθμός επιρροής τους στην κατηγοριοποίηση νέων προτύπων ή ακόμη και να εξαιρεθούν από την βάση γνώσης. Η διαδικασία αυτή έχει ως αποτέλεσμα την μείωση του αριθμού των προτύπων της βάσης και κατά συνέπεια ταχύτερη ανεύρεση των πλησιέστερων γειτόνων καθώς και βελτίωση της απόδοσης του αλγόριθμου [85].

□

Κεφάλαιο 3

Μεθοδολογίες ομαδοποίησης

Στο κεφάλαιο αυτό παρουσιάζονται δύο νέες μεθοδολογίες ομαδοποίησης. Η πρώτη βασίζεται στον συνδυασμό των αποτελεσμάτων της επαναληπτικής εκτέλεσης ενός βασικού αλγόριθμου [24]. Ενώ η δεύτερη βασίζεται στα αποτελέσματα της ομαδοποίησης μέσω αυτο-οργανούμενων χαρτών [56].

3.1 Ομαδοποίηση μέσω συγχώνευσης

Ο αλγόριθμος πολλαπλής συγχώνευσης αποτελείται από δύο φάσεις που εκτελούνται σειριακά: τη φάση της ομαδοποίησης που χρησιμοποιείται για να διαχωρίσει δεδομένα εισόδου σε ομάδες και τη φάση της συγχώνευσης κατά την οποία προκύπτει η τελική ομαδοποίηση των δεδομένων. Στην φάση της ομαδοποίησης καθορίζεται ο αρχικός αριθμός των ομάδων και ο αριθμός των επαναλήψεων της διαδικασίας. Κατά την φάση αυτή χρησιμοποιείται ένας αλγόριθμος ομαδοποίησης και ένα σχήμα ψηφοφορίας για να παραχθεί η ομαδοποίηση. Κατά την φάση της συγχώνευσης γίνεται επεξεργασία της ομαδοποίησης που προέκυψε και γειτονικές ομάδες συγχωνεύονται καταλήγοντας σε έναν βέλτιστο αριθμό ομάδων με βάση ορισμένα προκαθορισμένα κριτήρια. Η κεντρική ιδέα της μεθόδου αυτής είναι ότι με την χρήση ενός απλού βασικού αλγόριθμου ομαδοποίησης και με διαδοχικές επαναλήψεις του αλγόριθμου αυτού είναι δυνατόν να επιτευχθεί η ομαδοποίηση του συνόλου των δεδομένων εισόδου, συνδυάζοντας τα αποτελέσματα της κάθε επανάληψης έτσι ώστε να μην υπάρχει εξάρτηση από τις συνθήκες αρχικοποίησης του βασικού αλγόριθμου.

3.1.1 Φάση ομαδοποίησης

Κατά την φάση αυτή επαναλαμβάνεται η εκτέλεση ενός βασικού αλγόριθμου ομαδοποίησης για έναν προκαθορισμένο αριθμό επαναλήψεων με στόχο να πα-

ραχθούν διαφορετικές ομαδοποιήσεις, όσες και ο αριθμός των επαναλήψεων. Στη φάση αυτή χρησιμοποιήθηκε ο αλγόριθμος Fuzzy C-Means (FCM) [7]. Ο αλγόριθμος αυτός αναθέτει τα δεδομένα εισόδου σε ομάδες με ένα βαθμό συμμετοχής για το κάθε δεδομένο στην κάθε ομάδα. Ο αλγόριθμος χρησιμοποιεί έναν προκαθορισμένο αριθμό ομάδων και αναθέτει ένα κέντρο στην κάθε ομάδα. Στα επαναληπτικά βήματα που εκτελεί ανανεώνει τον ορισμό των κέντρων αυτών με στόχο την ελάττωση μίας συνάρτησης σφάλματος, η οποία ορίζεται μέσω της απόστασης του κάθε σημείου από το κάθε κέντρο σταθμισμένη με τον βαθμό συμμετοχής του σημείου στην ομάδα που αντιστοιχεί στο εκάστοτε κέντρο.

Η κάθε ομαδοποίηση που προκύπτει από τον αλγόριθμο αυτό είναι διαφορετική, εφόσον η αρχικοποίηση των κέντρων των ομάδων είναι διαφορετική. Προκύπτει επομένως το πρόβλημα της ταυτοποίησης της κάθε ομάδας σε σχέση με τις ομάδες που προέκυψαν από μία άλλη εκτέλεση. Για να αντιμετωπιστεί αυτό το πρόβλημα μετά το τέλος της κάθε ομαδοποίησης υπολογίζεται το ποσοστό των προτύπων που έχουν ανατεθεί σε κάθε ομάδα και κατά την προηγούμενη εκτέλεση αποτελούσαν πρότυπα μίας συγκεκριμένης ομάδας. Επομένως η ομάδα από την νέα ομαδοποίηση που κατέχει το μεγαλύτερο ποσοστό αντιστοιχίζεται με την ομάδα αυτή της προηγούμενης ομαδοποίησης. Εφόσον ο αριθμός των ομάδων που προκύπτει σε κάθε ομαδοποίηση είναι ο ίδιος, τότε δημιουργείται μία ένα προς ένα αντιστοιχία των ομάδων της εκάστοτε ομαδοποίησης με την προηγούμενη και αλυσιδωτά και με την πρώτη. Αυτό έχει ως αποτέλεσμα να είναι δυνατή η σύγκριση μεταξύ των ομάδων στις διαφορετικές ομαδοποιήσεις.

Μετά από αυτή την διαδικασία της ταυτοποίησης των ομάδων, εφαρμόζεται ένα σχήμα ψηφοφορίας για τα πρότυπα εισόδου. Κάθε πρότυπο εισόδου στις διαφορετικές ομαδοποιήσεις ανήκει σε κάποια από τις ομάδες. Ο σκοπός είναι να βρεθεί ένας βαθμός συμμετοχής με βάση το πόσες φορές το ίδιο πρότυπο αντιστοιχείται στην ίδια ομάδα. Αυτό επιτυγχάνεται με τον ορισμό ενός πίνακα ψηφοφορίας VT οποίος αποτελείται από N γραμμές και C στήλες, όπου N είναι ο αριθμός των προτύπων στο σύνολο των δεδομένων εισόδου και C ο αριθμός των ομάδων. Το κάθε πρότυπο αυξάνει το βαθμό συμμετοχής στην εκάστοτε ομάδα αν κατά την εκάστοτε ομαδοποίησης βρέθηκε ότι ανήκει σε αυτήν ενώ αντίστοιχα μειώνει τον βαθμό αυτό σε όλες τις υπόλοιπες ομάδες. Επομένως, κάθε πρότυπο ψηφίζει θετικά ή αρνητικά τις ομάδες τόσες φορές όσες και οι διαφορετικές ομαδοποιήσεις που προέκυψαν από τις εκτελέσεις του βασικού αλγόριθμου.

Ο βαθμός συμμετοχής ενός προτύπου $\mathbf{x}_i$ στην ομάδα j καθορίζεται τελικά από την τιμή $VT(i, j)$ και η τελική ομάδα $C^{\mathbf{x}}_{max}$ στην οποία ανατίθεται το πρότυπο αυτό είναι αυτή με στην οποία εμφανίζει τον μέγιστο βαθμό συμμετοχής.

$$C_{\max}^{\mathbf{x}_i} = \arg\max_j \left(VT(i,j)\right), \;\; j = 1, 2, ..., C \tag{3.1}$$

Χρησιμοποιώντας τον πίνακα ψηφοφορίας και την σχέση μεταξύ των προτύπων της μίας ομάδας με όλες τις υπόλοιπες, ορίζεται ο πίνακας NRT (διάστασης C επί C), έτσι ώστε η τιμή $NRT(m, j)$ να αντιπροσωπεύει τον βαθμό συγγένειας μεταξύ των ομάδων m και j.

$$NRT(m,j) = \sum_{p=1}^{N} (VT(p,j) I(C_{\max}^{p} = 1)), \;\; m = 1, 2, ..., C, \;\; j = 1, 2, ..., C, \;\; j \neq i \tag{3.2}$$

όπου $I(z)$ είναι μία συνάρτηση που αποτιμά την λογική έκφραση z και επιστρέφει 1 εφόσον η έκφραση αληθεύει και 0 στην αντίθετη περίπτωση. Ο αριθμός C των ομάδων όπως και προαναφέρθηκε είναι προκαθορισμένος κατά την εκτέλεση του αλγόριθμου. Ο στόχος όμως στην φάση αυτή δεν είναι να βρεθεί ο βέλτιστος θεωρητικά αριθμός των ομάδων που υπάρχουν στο σύνολο των δεδομένων, άρα επιλέγεται τιμή του της παραμέτρου C μεγαλύτερη του 50 έτσι ώστε η ομαδοποίηση που θα προκύψει να δώσει πολλές και μικρές ομάδες, σε σχέση με το πλήθος των στοιχείων που περιέχουν.

3.1.2 Φάση συγχώνευσης

Κάνοντας χρήση του βαθμού συγγένειας που υπολογίζεται με την συνάρτηση (3.2), εκτελείται η φάση της συγχώνευσης. Η διαδικασία αυτή ξεκινά με τον προκαθορισμένο από την προηγούμενη φάση αριθμό ομάδων C και συγχωνεύει τις ομάδες με τον μεγαλύτερο βαθμό συγγένειας. Πιο συγκεκριμένα η διαδικασία εντοπίζει στον πίνακα NRT τις δύο ομάδες ($C1$ και $C2$) που πληρούν τις εξής συνθήκες: α)για την κάθε ομάδα η πλησιέστερη ομάδα είναι η άλλη και β)οι δύο ομάδες είναι οι πλησιέστερες στο σύνολο των C ομάδων. Το επόμενο βήμα είναι η συγχώνευση των δύο αυτών ομάδων και ο επαναπροσδιορισμός του πίνακα ψηφοφορίας VT με την πρόσθεση των ψήφων της ομάδας $C1$ στους ψήφους της ομάδας $C2$ όπως φαίνεται και στην εξίσωση (3.3).

$$VT(i, C1') = VT(i, C1') + VT(i, C2'), \;\; i = 1, 2, ..., N \tag{3.3}$$

όπου $C1' = min(C1, C2)$ και $C2' = max(C1, C2)$. Ο νέος πίνακας NRT υπολογίζεται με μία ομάδα λιγότερο, αφού η ομάδα $C2'$ συγχωνεύτηκε με την ομάδα $C1'$ και η διαδικασία επαναλαμβάνεται από την αρχή έως την πλήρωση κάποιου κριτηρίου τερματισμού. Το κριτήριο τερματισμού που χρησιμοποιείται

για να τερματίσει την διαδικασία συγχώνευσης των ομάδων είναι ο συνδυασμός των μέτρων RMSSTD (Root Mean Square Standard Deviation) και RS (R-Squared) [65]. Πιο συγκεκριμένα, για ένα σύνολο προτύπων $\mathbf{x}_i, i = 1, ..., N$ ορίζεται η τιμή SS ως το παρακάτω άθροισμα τετραγώνων.

$$SS = \sum_{k=1}^{d} \sum_{i=1}^{N} (x_{ik} - \overline{x_k})^2 \tag{3.4}$$

Όπου x_{ik} είναι η τιμή του χαρακτηριστικού k του προτύπου $\mathbf{x}_i$ ενώ $\overline{x_k}$ είναι η μέση τιμή των τιμών του χαρακτηριστικού k για όλα τα πρότυπα του συνόλου και d είναι το πλήθος των χαρακτηριστικών, δηλαδή η διάσταση του χώρου των δεδομένων εισόδου. Με βάση τον ορισμό της τιμής SS ορίζονται οι τιμές:

1. SS_{wj} για το σύνολο των προτύπων που ανήκουν στην ομάδα j.
2. SS_t για όλα τα πρότυπα του συνόλου των δεδομένων εισόδου.

Ο δείκτης RS ορίζεται ως:

$$RS = \frac{SS_t - \sum_{j=1}^{C} SS_{wj}}{SS_t} \tag{3.5}$$

όπου C είναι το πλήθος των ομάδων. Ο δείκτης αυτός κυμαίνεται μεταξύ των τιμών 0 και 1 και αξιολογεί την ανομοιογένεια των ομάδων, η οποία είναι ανάλογη της τιμής του δείκτη.

Ο δείκτης $RMSSTD$ ορίζεται ως:

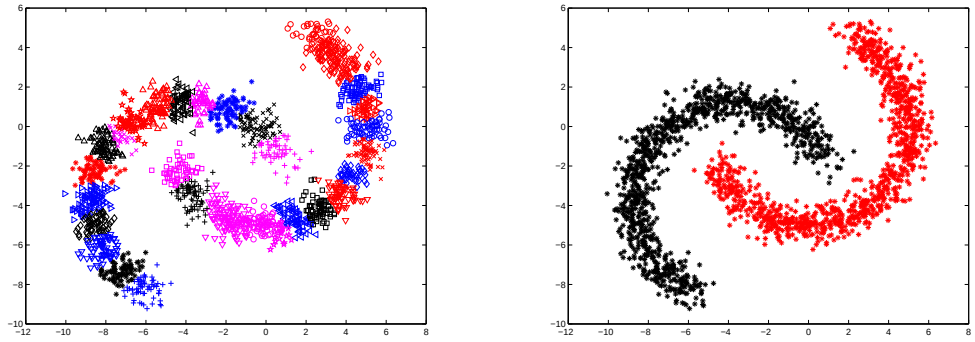
$$RMSSTD = \sqrt{\frac{\sum_{j=1}^{C} SS_{wj}}{\sum_{j=1}^{C} d(N_j - 1)}} \tag{3.6}$$

όπου N_j είναι το πλήθος των προτύπων που ανήκουν στην ομάδα j. Ο δείκτης αυτός, αντίθετα με τον προηγούμενο, αξιολογεί την ομοιογένεια των ομάδων. Οι δύο αυτοί δείκτες μεταβάλλονται ομαλά καθώς επαναυπολογίζονται μετά από κάθε συγχώνευση. Εάν μετά από μία συγχώνευση παρατηρηθεί μία μεταβολή των δεικτών σημαντικά μεγαλύτερη συγκριτικά με τις προηγούμενες τότε αυτό αποτελεί και ένδειξη για την διακοπή της διαδικασίας των συγχωνεύσεων.

3.1.3 Πειραματική αξιολόγηση

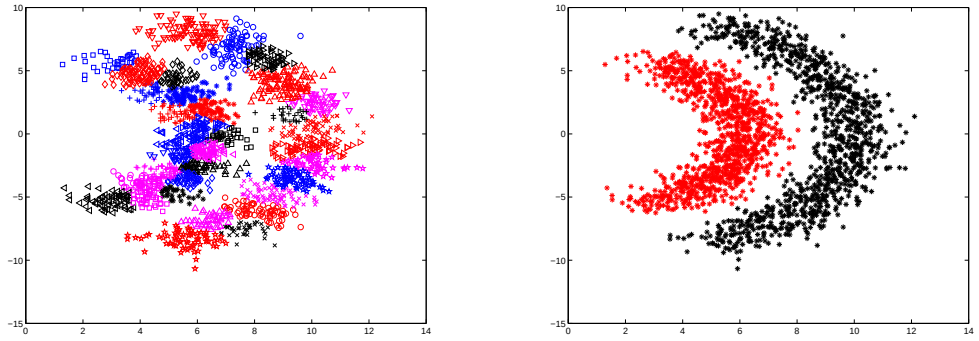
Η πειραματική αξιολόγηση έγινε με την χρήση δεδομένων δύο διαστάσεων έτσι ώστε να είναι δυνατή η οπτικοποίηση των αποτελεσμάτων. Παρακάτω παρου-

σιάζονται τα αποτελέσματα από δύο σύνολα δεδομένων που έχουν ως κύριο χαρακτηριστικό ότι στοιχεία τους σχηματίζουν ομάδες μη κυκλικού σχήματος (σχήματα 3.1, 3.2). Όπως φαίνεται στα σχήματα 3.1.α και 3.2.α, στο τέλος της πρώτης φάσης έχουν σχηματιστεί πολλές μικρές ομάδες. Στα σχήματα 3.1.β και 3.2.β, μετά την φάση της συγχώνευσης οι μικρότερες ομάδες έχουν συγχωνευθεί σχηματίζοντας τις δύο μεγαλύτερες ομάδες που είναι και το επιθυμητό αποτέλεσμα.



(α) (β)

Σχήμα 3.1: *Ομαδοποίηση του συνόλου δεδομένων Banana. (α) Σχηματισμός ομάδων μετά το τέλος της πρώτης φάσης. (β) Τελική ομαδοποίηση μετά και την φάση της συγχώνευσης.*



(α) (β)

Σχήμα 3.2: *Ομαδοποίηση του συνόλου δεδομένων Lith. (α) Σχηματισμός ομάδων μετά το τέλος της πρώτης φάσης. (β) Τελική ομαδοποίηση μετά και την φάση της συγχώνευσης.*

3.2 Ομαδοποίηση με χρήση αυτο-οργανούμενων χαρτών

Στο πλαίσιο της έρευνας για ανάπτυξη υβριδικών συστημάτων διερευνήθηκε η δυνατότητα χρήσης της επεξεργασίας δεδομένων με αυτο-οργανούμενους χάρτες (Self Organizing Maps - SOM), με τελικό στόχο την ενσωμάτωση τους σε ένα υβριδικό μοντέλο. Οι αυτο-οργανούμενοι χάρτες [41],[42], χρησιμοποιώντας μάθηση χωρίς επίβλεψη, πετυχαίνουν μείωση της διάστασης των δεδομένων σε έναν περιορισμένο αριθμό νευρώνων. Συγχρόνως, καταφέρνουν να διατηρούν την τοπολογία του χώρου δεδομένων και, γι' αυτό τον λόγο, θεωρούνται κατάλληλοι για ανάλυση προβλημάτων με δεδομένα πολλών διαστάσεων.

3.2.1 Περιγραφή μεθοδολογίας ομαδοποίησης

Στο πλαίσιο αυτό, αναπτύχθηκε μία μεθοδολογία με σκοπό την αυτόματη εξαγωγή ομάδων των δεδομένων εισόδου από έναν εκπαιδευμένο χάρτη SOM, καθώς και αξιολόγηση των χαρακτηριστικών των δεδομένων εισόδου ως προς την επίδραση που έχουν στη δημιουργία των ομάδων αυτών. Αφού εκπαιδευθεί ο χάρτης με τα δεδομένα εισόδου, στη συνέχεια εκτελείται ένας ιεραρχικός αλγόριθμος ομαδοποίησης λαμβάνοντας ως δεδομένα εισόδου τα χαρακτηριστικά των νευρώνων του χάρτη. Επίσης ο αλγόριθμος εκμεταλλεύεται και την πληροφορία της γειτνίασης των νευρώνων πάνω στον χάρτη καθώς και της πυκνότητας των αρχικών δεδομένων εισόδου στους νευρώνες του εκπαιδευμένου χάρτη. Δηλαδή, υπάρχει η προϋπόθεση ότι για να μπορούν δύο νευρώνες να τοποθετηθούν στην ίδια ομάδα πρέπει να υπάρχει μεταξύ τους σχέση γειτνίασης, όπως και να πληρούν και οι δύο κάποια κριτήρια ως προς την πυκνότητα των δεδομένων εισόδου. Εφαρμόζοντας την παραπάνω διαδικασία προκύπτει μια ομαδοποίηση των νευρώνων του χάρτη.

Ο δεύτερος στόχος είναι η αξιολόγηση των χαρακτηριστικών των δεδομένων εισόδου. Εφόσον έχει πραγματοποιηθεί η αρχική ομαδοποίηση των δεδομένων εισόδου με την εκπαίδευση του χάρτη SOM και αφού οι νευρώνες του χάρτη περιέχουν συμπιεσμένη την πληροφορία των δεδομένων εισόδου και, επίσης, τα χαρακτηριστικά των νευρώνων ανταποκρίνονται και στα χαρακτηριστικά των δεδομένων εισόδου, τότε για να επιτευχθεί οικονομία χρόνου και υπολογιστικών πόρων μπορεί να πραγματοποιηθεί η ανάλυση των χαρακτηριστικών των δεδομένων εισόδου αναλύοντας τα χαρακτηριστικά των νευρώνων. Αυτό επιτυγχάνεται ως εξής: η παραπάνω διαδικασία ομαδοποίησης επαναλαμβάνεται τόσες φορές όσες και τα χαρακτηριστικά των δεδομένων εισόδου. Σε κάθε επανάληψη της

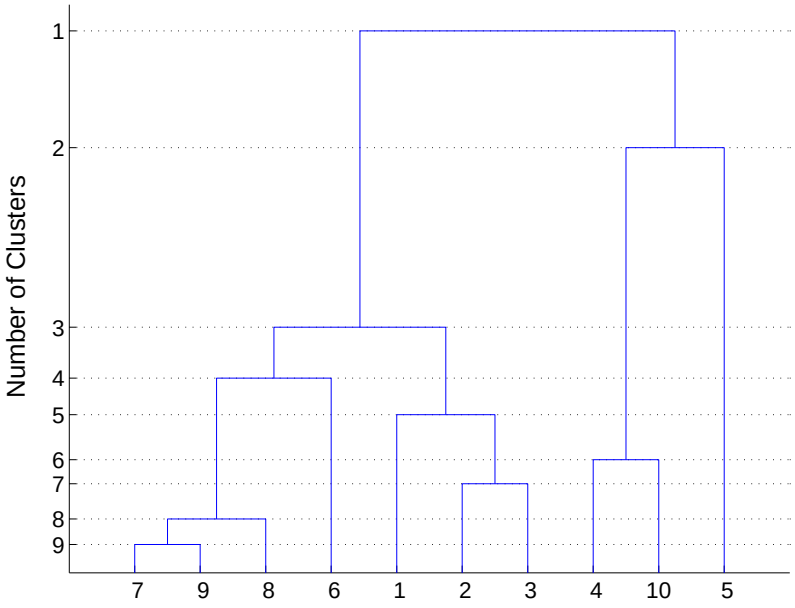
διαδικασίας ένα από τα χαρακτηριστικά ευνοείται απέναντι των άλλων. Αυτό έχει ως αποτέλεσμα σε κάθε επανάληψη να προκύπτει μία νέα ομαδοποίηση, στην οποία το αποτέλεσμα έχει επηρεαστεί από την εύνοια απέναντι στο εκάστοτε χαρακτηριστικό. Συγκρίνοντας κάθε φορά το αποτέλεσμα της ομαδοποίησης με την ομαδοποίηση που προέκυψε από την εφαρμογή του αλγόριθμου στον αρχικά εκπαιδευμένο χάρτη, μπορούν να εξαχθούν συμπεράσματα για την επίδραση του εκάστοτε χαρακτηριστικού εισόδου σε σχέση με τις ομάδες που προέκυψαν από την αρχική ομαδοποίηση. Συνοπτικά τα βήματα που ακολουθούνται είναι τα εξής:

- Εκπαίδευση ενός χάρτη SOM με τα δεδομένα εισόδου του προβλήματος.
- Ομαδοποίηση των νευρώνων του χάρτη.
- Επανάληψη της ομαδοποίησης τόσες φορές όσα και τα χαρακτηριστικά εισόδου του προβλήματος.
- Σύγκριση της κάθε ομαδοποίησης με την αρχική ομαδοποίηση για αξιολόγηση του κάθε χαρακτηριστικού εισόδου.

Σαν συμπέρασμα μπορεί να διατυπωθεί ότι με την χρήση της παραπάνω μεθοδολογίας επιτυγχάνεται η ομαδοποίηση των δεδομένων εισόδου καθώς και η αξιολόγηση των χαρακτηριστικών τους ως προς τις ομάδες που προέκυψαν, χρησιμοποιώντας τα πλεονεκτήματα δύο διαφορετικών μεθοδολογιών, των αυτοοργανούμενων χαρτών καθώς και των ιεραρχικών αλγορίθμων ομαδοποίησης. Τα δύο τελευταία στάδια της μεθοδολογίας, δηλαδή η διαδικασία αξιολόγησης των χαρακτηριστικών περιγράφεται λεπτομερώς στο επόμενο κεφάλαιο (παράγραφος 4.2.1) όπου συμμετέχει στην δημιουργία ενός υβριδικού συστήματος κατηγοριοποίησης.

Η διαδικασία της ομαδοποίησης και της αξιολόγησης των χαρακτηριστικών πιο αναλυτικά γίνεται ως εξής. Χρησιμοποιώντας τα διανύσματα των νευρώνων του πλέγματος ως δεδομένα εισόδου εκτελείται ένας ιεραρχικός συγκεντρωτικός αλγόριθμος ομαδοποίησης. Ο αλγόριθμος αυτός αρχικοποιεί κάθε δεδομένο εισόδου ως διακριτή ομάδα. Σε κάθε βήμα εκτέλεσης του οι δυο πλησιέστερες μεταξύ τους ομάδες επιλέγονται για να ενωθούν και να σχηματίσουν μία ενιαία ομάδα. Αν δεν χρησιμοποιηθεί κάποιο κριτήριο τερματισμού των συνενώσεων τότε όλα τα δεδομένα εισόδου θα ενταχθούν σε μία, ενιαία ομάδα. Στο σχήμα 3.3 φαίνεται ενδεικτικά το δένδρο των συνενώσεων από δέκα ομάδες σε μία.

Τα παραμετροποιήσιμα χαρακτηριστικά του αλγόριθμου είναι: α) ο ορισμός της απόστασης μεταξύ δύο ομάδων έτσι ώστε να δυνατόν να βρεθούν οι δύο



Σχήμα 3.3: *Δένδρο συγχωνεύσεων.*

πλησιέστερες και β) το κριτήριο τερματισμού της διαδικασίας πριν σχηματιστεί μία και μοναδική ομάδα.

Έστω δύο ομάδες A, B και N_A, N_B ο αριθμός των δεδομένων της κάθε ομάδας αντίστοιχα. Αν $\mathbf{w}_i$ και $\mathbf{w}_j$ είναι πρότυπα που ανήκουν στις ομάδες A και B αντίστοιχα τότε η απόσταση $D_{A,B}$ μεταξύ των ομάδων μπορεί να οριστεί με διάφορους τρόπους όπως φαίνεται και στον πίνακα 3.1.

Πίνακας 3.1: *Μετρικές αποστάσεων μεταξύ ομάδων*

Ελάχιστη απόσταση	$D_{A,B} = \min_{i,j}(\|\mathbf{w}_i - \mathbf{w}_j\|)$
Μέγιστη απόσταση	$D_{A,B} = \max_{i,j}(\|\mathbf{w}_i - \mathbf{w}_j\|)$
Μέση απόσταση	$D_{A,B} = \frac{\sum_{i,j}\|\mathbf{w}_i - \mathbf{w}_j\|}{N_A \cdot N_B}$
Απόσταση κέντρων	$D_{A,B} = \|\mathbf{c}_A - \mathbf{c}_B\|$

$$i = 1, 2, ..., N_A, \quad j = 1, 2, ..., N_B$$

Στην προκειμένη περίπτωση, ο αλγόριθμος χρησιμοποιεί σαν δεδομένα εισόδου τα διανύσματα των νευρώνων του πλέγματος. Αυτό δίνει την δυνατότητα να χρησιμοποιηθεί και η πληροφορία της θέσης των νευρώνων επάνω στο πλέγμα. Στην διαδικασία χρησιμοποιούνται δύο επιπλέον κριτήρια με στόχο να αξιοποι-

ηθεί η παραπάνω πληροφορία αλλά και να προκύψει και μία ομαδοποίηση που θα είναι συμβατή με τον εκπαιδευμένο χάρτη.

Στον υπολογισμό της απόστασης χρησιμοποιείται η μετρική της ελάχιστης απόστασης και από τα διανύσματα των δύο ομάδων μεταξύ των οποίων υπολογίζεται η απόσταση συμμετέχουν μόνο τα ζεύγη εκείνα των οποίων οι νευρώνες συνδέονται στο πλέγμα. Επομένως η απόσταση $D_{A,B}$ υπολογίζεται ως:

$$D_{A,B} = \min_{i,j}(\|\mathbf{w}_i - \mathbf{w}_j\|), \quad i = 1, 2, ..., N_A, \quad j = 1, 2, ..., N_B, \\ if \quad and \quad only \quad if \quad i, j \quad adjacent \quad units \tag{3.7}$$

Με τη χρήση αυτή της συνθήκης, προφανώς οι ομάδες που δεν έχουν συνδεόμενους νευρώνες δεν μπορούν να ενωθούν και αυτό εξασφαλίζει ότι σε κάθε ομάδα που θα δημιουργηθεί, οι νευρώνες που θα περιλαμβάνει θα καταλαμβάνουν στο χάρτη μια ενιαία και αδιαίρετη περιοχή.

Μετά την εκπαίδευση του χάρτη από το αρχικό σύνολο δεδομένων, το σύνολο αυτό παρουσιάζεται εκ νέου στο χάρτη όχι όμως για εκπαίδευση και προσαρμογή των νευρώνων αλλά για να αντιστοιχιστεί στο κάθε πρότυπο εισόδου ο τελικός νευρώνας νικητής. Αυτό έχει ως αποτέλεσμα κάθε νευρώνας του χάρτη να συγκεντρώνει έναν αριθμό από τα πρότυπα εισόδου. Οι νευρώνες που δεν συγκεντρώνουν κανένα πρότυπο εισόδου ή συγκεντρώνουν ένα μικρό ποσοστό του συνόλου των προτύπων μπορούν εξαιρεθούν από την μετέπειτα διαδικασία ομαδοποίησης καθώς θεωρείται ότι τα διανύσματα των νευρώνων αυτών αντιστοιχούν σε περιοχές του χώρου εισόδου στις οποίες η συγκέντρωση των προτύπων είναι χαμηλή και ως εκ τούτου αποτελούν διαχωριστικές περιοχές μεταξύ των ομάδων που υπάρχουν στα πρότυπα εισόδου.

Ο συνδυασμός των παραπάνω κριτηρίων περιορίζει τις πιθανές συνενώσεις μεταξύ των ομάδων που σχηματίζονται κατά την εκτέλεση του ιεραρχικού αλγόριθμου. Κατά το στάδιο της συνένωσης δύο ομάδων εξετάζεται αν οι προς συνένωση ομάδες πληρούν άλλη μία συνθήκη. Για κάθε ομάδα ορίζεται ένα εσωτερικό μέτρο συνεκτικότητας. Στον πίνακα 3.2 αναφέρονται μετρικές που μπορούν να χρησιμοποιηθούν σαν τέτοια μέτρα.

Στην παρούσα μεθοδολογία οι δύο πρώτες μετρικές τροποποιούνται έτσι ώστε οι αποστάσεις να υπολογίζονται μόνο για τους νευρώνες που είναι μεταξύ τους συνδεδεμένοι στο πλέγμα. Έχοντας υπολογίσει το μέτρο συνεκτικότητας της κάθε ομάδας, εξετάζεται η παρακάτω συνθήκη. Η ομάδα που θα προκύψει από την συγχώνευση να μην έχει σημαντικά μειωμένη συνεκτικότητα σε σχέση με τις δύο ομάδες που την δημιούργησαν. Η συνθήκη αυτή χρησιμοποιείται έτσι

Πίνακας 3.2: *Μέτρα συνεκτικότητας ομάδων*

Ολική Μέση απόσταση	$IC_A = \frac{\sum_{i,j} \|\mathbf{w}_i - \mathbf{w}_j\|}{N_A \cdot (N_A - 1)}$
Μέση Ελάχιστη απόσταση	$NNIC_k = \frac{\sum_{i=1}^{N_k} \min_{ij, j=1..N_k, (i \neq j)} \{\|\mathbf{m}_i - \mathbf{m}_j\|\}}{N_k}$
Μέση Απόσταση από το κέντρο	$IC_A = \frac{\sum_i \|\mathbf{w}_i - \mathbf{c}_A\|}{N_A}$

$$i = 1, 2, ..., N_A, \quad j = 1, 2, ..., N_B$$

ώστε δύο ομάδες που παρόλο η απόσταση μεταξύ τους είναι η μικρότερη σε σχέση με τις υπόλοιπες να μην συγχωνεύονται εφόσον εμφανίζουν ένα βαθμό συνεκτικότητας ο οποίος μετά την συγχώνευση θα αλλοιωθεί σημαντικά. Έτσι εξασφαλίζεται η διατήρηση συμπαγών διακριτών ομάδων. Επίσης εξυπηρετεί και την διαδικασία τερματισμού των συνενώσεων πριν αυτές οδηγήσουν σε μία και μοναδική ομάδα.

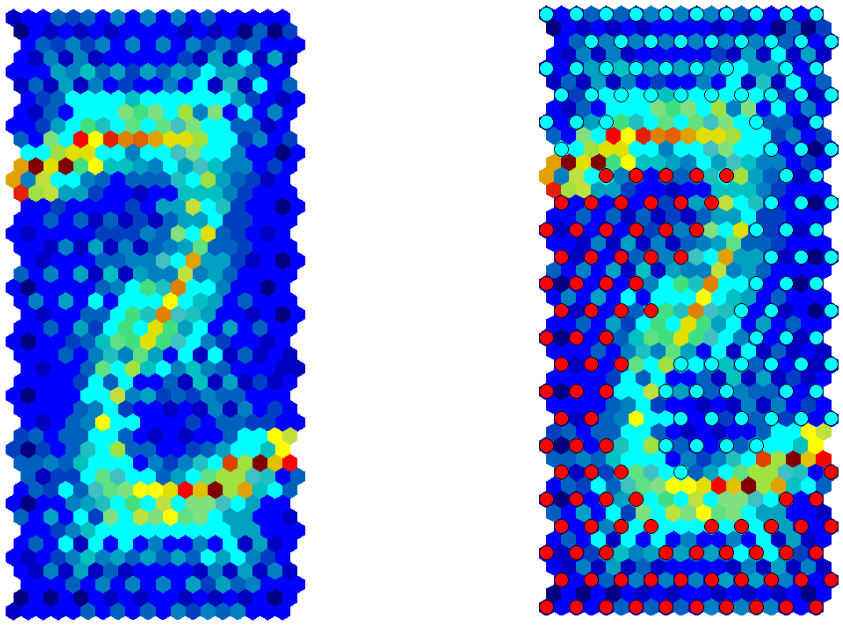
Εφόσον χρησιμοποιείται σαν μέτρο συνεκτικότητας η μέση ελάχιστη απόσταση και σαν μέτρο απόστασης μεταξύ δύο ομάδων η ελάχιστη απόσταση τότε μία πιο απλουστευμένη υλοποίηση του παραπάνω κριτηρίου θα ήταν η σύγκριση της απόστασης μεταξύ των δύο ομάδων με τον βαθμό συνεκτικότητάς τους όπως ορίζεται στην παρακάτω εξίσωση:

$$D_{A,B} \leq L \cdot \frac{IC_A + IC_B}{2} \tag{3.8}$$

όπου η παράμετρος L εκφράζει τον βαθμό ανοχής που επιτρέπεται η απόσταση μεταξύ των δύο ομάδων να υπερβαίνει τον μέσο όρο των βαθμών συνεκτικότητάς τους. Η διαδικασία συνένωσης περιορίζεται από όλες τις παραπάνω συνθήκες, οι οποίες όμως δεν εξασφαλίζουν ότι η διαδικασία θα τερματιστεί έχοντας καταλήξει στο ιδανικό αριθμό των ομάδων. Για το λόγο αυτό χρησιμοποιούνται δύο επιπλέον μέτρα. Ο συνδυασμός των μέτρων RMSSTD (Root Mean Square Standard Deviation) και RS (R- Squared) [65] επιτρέπει τον όσον το δυνατό καλύτερο προσδιορισμό του βέλτιστου αριθμού των ομάδων.

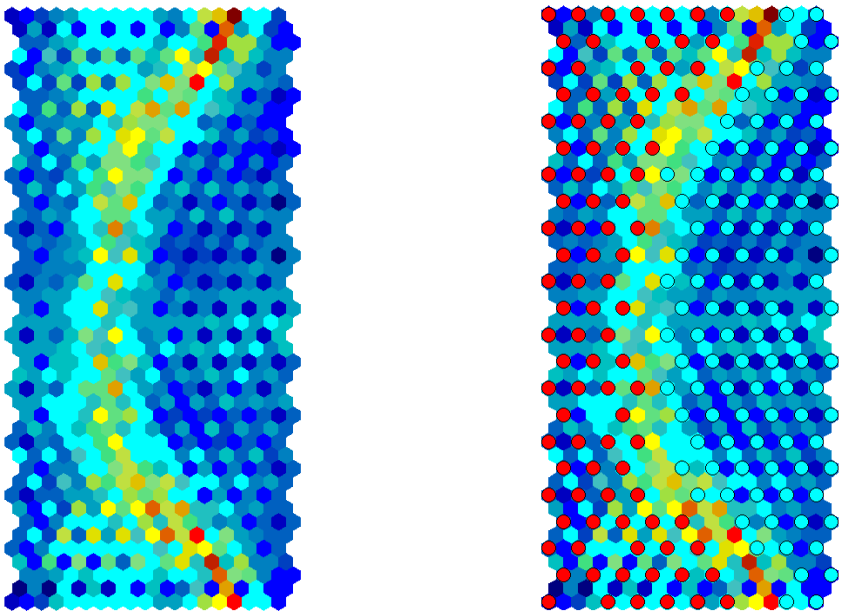
3.2.2 Πειραματική αξιολόγηση

Για την επαλήθευση της ορθής λειτουργίας της μεθοδολογίας παρατίθενται τα αποτελέσματα της εφαρμογής της στα σύνολα δεδομένων που χρησιμοποιήθηκαν και στην παράγραφο 3.1.3. Στα σχήματα 3.4.α και 3.5.α παρουσιάζονται οι διατάξεις U-matrix των αυτο-οργανούμενων χαρτών που εκπαιδεύτηκαν με τα παραπάνω σύνολα δεδομένων. Ο χρωματισμός του χάρτη είναι ανάλογος της Ευκλείδειας απόστασης μεταξύ των συνδεδεμένων νευρώνων του πλέγματος. Η απόσταση υπολογίζεται χρησιμοποιώντας τα διανύσματα των νευρώνων. Οι αποχρώσεις του μπλε χρώματος αντιστοιχούν στις μικρότερες αποστάσεις ενώ αντίθετα οι αποχρώσεις του κόκκινου αντιστοιχούν στις μεγαλύτερες αποστάσεις. Οι ενδιάμεσοι χρωματισμοί αντιστοιχούν σε ανάλογες αποστάσεις. Και στα δυο σχήματα είναι εμφανές ότι το πλέγμα έχει διαχωρίσει τις δυο ομάδες των δεδομένων σε δυο διακριτές περιοχές του πλέγματος με μπλε χρωματισμούς που διαχωρίζονται από μια επιφάνεια κόκκινων αποχρώσεων. Στα σχήματα 3.4.β και 3.5.β, οι νευρώνες έχουν ομαδοποιηθεί και στις δύο περιπτώσεις σε δύο διαφορετικές ομάδες, που επισημαίνονται με κύκλους διαφορετικού χρώματος. Η μεθοδολογία έχει καταφέρει να διαχωρίσει σωστά τις περιοχές του χάρτη που αντιστοιχούν στις διαφορετικές ομάδες των δεδομένων εισόδου.



(α) (β)

Σχήμα 3.4: *Ομαδοποίηση του συνόλου δεδομένων Banana. (α) Διάταξη U-matrix του χάρτη SOM. (β) Ομαδοποίηση των νευρώνων του χάρτη.*



(α) (β)

Σχήμα 3.5: *Ομαδοποίηση του συνόλου δεδομένων Lith. (α) Διάταξη U-matrix του χάρτη SOM. (β) Ομαδοποίηση των νευρώνων του χάρτη.*

3.3 Συζήτηση - Συμπεράσματα

Στο κεφάλαιο αυτό παρουσιάστηκαν δύο μεθοδολογίες ομαδοποίησης, οι οποίες φαινομενικά είναι διαφορετικές. Όμως βασίζονται σε μία κοινή προσέγγιση. Το πρώτο στάδιο και των δύο μεθοδολογιών είναι μία διαδικασία ανεύρεσης ενός πολυπληθούς συνόλου ομάδων των προτύπων που αποτελούν το σύνολο των δεδομένων εισόδου. Στην συνέχεια μέσω συγχωνεύσεων των ομάδων προκύπτει ένας μικρότερος και κατά προσέγγιση βέλτιστος αριθμός ομάδων. Η πρώτη μεθοδολογία προσπαθεί να σχηματίσει το πολυπληθές σύνολο των αρχικών ομάδων μέσω της επαναληπτικής εκτέλεσης ενός σχετικά απλού αλγόριθμου όπως ο αλγόριθμος Fuzzy C-Means και συγχρόνως να τροφοδοτήσει το επόμενο στάδιο με την πληροφορία της συγγένειας μεταξύ των ομάδων, η οποία θα αποτελέσει και το κριτήριο για την περαιτέρω συγχώνευση των ομάδων.

Η μεθοδολογία αυτή αυξάνει τον χρόνο ομαδοποίησης σε σχέση με τον αλγόριθμό Fuzzy C-Means ή όποιον άλλον αλγόριθμο χρησιμοποιηθεί στην θέση του, καθώς τον επαναλαμβάνει έτσι ώστε να συνδυάσει τα αποτελέσματα. Επίσης εισάγεται επιπλέον χρονική καθυστέρηση με την επεξεργασία των αποτελεσμάτων των πολλαπλών εκτελέσεων. Η μεθοδολογία όμως παρουσιάζει την δυνατότητα να εντοπίζει ομάδες δεδομένων μη σφαιρικού σχήματος. Η δυνατότητα αυτή απουσιάζει από τον αλγόριθμο Fuzzy C-Means, οπότε η χρονική καθυστέρηση

που εισάγεται είναι το αντίτιμο για την ριζική βελτίωση αυτού του απλού αλγορίθμου.

Η δεύτερη μεθοδολογία χρησιμοποιεί στο πρώτο στάδιο ένα θεωρητικά πιο σύνθετο μοντέλο, αυτό των αυτο-οργανούμενων χαρτών. Το μοντέλο αυτό αποτελεί και γενίκευση των αλγορίθμων ομαδοποίησης με χρήση κεντροειδών πρωτοτύπων. Σε αυτή την περίπτωση το σύνολο των αρχικών ομάδων μπορεί να θεωρηθεί ότι αποτελούν τα σύνολα των πρότυπων που αντιστοιχούν σε κάθε νευρώνα του πλέγματος του εκπαιδευμένου χάρτη. Η επανάληψη της εκτέλεσης της αρχικής ομαδοποίησης δεν είναι απαραίτητη καθώς το μοντέλο των αυτο-οργανούμενων χαρτών δεν επηρεάζεται τόσο από τις αρχικές τιμές των νευρώνων του πλέγματος. Επίσης η συγγένεια των ομάδων προκύπτει εδώ σε κάποιο βαθμό από την τοπολογική σχέση μεταξύ των αντίστοιχων νευρώνων επάνω στο πλέγμα.

Στην περίπτωση της μεθοδολογίας αυτής, η χρονική καθυστέρηση που εισάγεται με την επεξεργασία των αποτελεσμάτων του αυτο-οργανούμενου χάρτη είναι αμελητέα, καθώς το σύνολο εισόδου του σταδίου αυτού, το οποίο είναι το σύνολο των διανυσμάτων των νευρώνων, έχει συνήθως πλήθος πολύ μικρότερο από το σύνολο των δεδομένων.

Η πειραματική αξιολόγηση και των δύο μεθοδολογιών έδειξε ότι παρουσιάζουν την ικανότητα να διαχωρίζουν με επιτυχία σύνολα δεδομένων που χαρακτηρίζονται από μη κυκλικούς σχηματισμούς και στα οποία αρκετοί αλγόριθμοι ομαδοποίησης εμφανίζουν προβληματική συμπεριφορά.

□

Κεφάλαιο 4

Υβριδικό νευρο-ασαφές μοντέλο κατηγοριοποίησης

4.1 Εισαγωγή

Με σκοπό την αξιολόγηση της δυνατότητας συμμετοχής διάφορων μοντέλων εξόρυξης γνώσης αναπτύχθηκε ένα πρωτότυπο υβριδικό σύστημα [56]. Το μοντέλο βασίζεται σε δύο ευρέως γνωστά μοντέλα υπολογιστικής νοημοσύνης, τους αυτο-οργανούμενους χάρτες και τα νευρο-ασαφή συστήματα. Ο στόχος ήταν η δημιουργία ενός συστήματος κατηγοριοποίησης, το οποίο θα εκμεταλλεύεται τα πλεονεκτήματα των νευρο-ασαφών συστημάτων κατηγοριοποίησης και, συγχρόνως, για να βελτιώσει ακόμα περισσότερο την απόδοση τέτοιων συστημάτων, θα χρησιμοποιεί μετα-δεδομένα που θα προκύπτουν από την επεξεργασία των δεδομένων με την χρήση αυτο-οργανούμενων χαρτών.

Τα νευρο-ασαφή συστήματα κατηγοριοποίησης είναι υβριδικά συστήματα που αποτελούνται από συνδυασμό μοντέλων ασαφούς λογικής και νευρωνικών δικτύων. Η χρήση ασαφούς λογικής εγγυάται την καλύτερη επεξεργασία ποσοτικών χαρακτηριστικών και επιτρέπει την εισαγωγή εννοιών που βρίσκονται εγγύτερα στον τρόπο λειτουργίας της ανθρώπινης λογικής. Ο συνδυασμός τους με τα νευρωνικά δίκτυα δίνει την δυνατότητα αυτόματης εξαγωγής ασαφών κανόνων από τα δεδομένα εισόδου με σκοπό την κατηγοριοποίηση των δεδομένων. Η απόδοση όμως αυτών των συστημάτων επηρεάζεται σε μεγάλο βαθμό από την αρχικοποίηση των ελεύθερων παραμέτρων τους. Επίσης, όταν τα συστήματα αυτά επεξεργάζονται δεδομένα υψηλών διαστάσεων, ο χρόνος καθώς και οι υπολογιστικοί πόροι που απαιτούνται για την εκπαίδευση τους αυξάνονται δραματικά.

Έχει αποδειχθεί ότι, με εξαγωγή γνώσης από τα δεδομένα εισόδου και με

χρήση της γνώσης αυτής για την αρχικοποίηση νευρο-ασαφών συστημάτων, βελτιώνεται σημαντικά η απόδοση των τελευταίων [25]. Το πρωτότυπο υβριδικό σύστημα που αναπτύχθηκε έχει ως στόχο την εξαγωγή γνώσης από τα δεδομένα με μεθόδους μη επιβλεπόμενης μάθησης και στην συνέχεια ενσωμάτωση της γνώσης αυτής σε ένα νευρο-ασαφές σύστημα προσαρμοζόμενης δομής.

4.2 Περιγραφή λειτουργίας συστήματος

Το σύστημα λειτουργεί εκτελώντας τα παρακάτω βήματα :

1. Εκπαίδευση χάρτη SOM.
2. Ομαδοποίηση των νευρώνων του χάρτη.
3. Αξιολόγηση των χαρακτηριστικών εισόδου.
   (α) Επανάληψη της ομαδοποίησης των νευρώνων για κάθε χαρακτηριστικό εισόδου.
   (β) Συνδυασμός των αποτελεσμάτων με τα αποτελέσματα του αρχικού χάρτη.
4. Αρχικοποίηση του νευρο-ασαφούς μοντέλου ταξινόμησης.
5. Επιβλεπόμενη εκπαίδευση νευρο-ασαφούς μοντέλου.
   (α) Υπολογισμός του ρυθμού μάθησης με βάση την προηγούμενη αξιολόγηση των χαρακτηριστικών εισόδου.
   (β) Προσαρμογή σε πραγματικό χρόνο της τοπολογίας του νευρο-ασαφούς μοντέλου.

Τα δύο πρώτα στάδια της λειτουργίας του συστήματος χρησιμοποιούν την μεθοδολογία που περιγράφηκε στο προηγούμενο κεφάλαιο. Στο τρίτο στάδιο εφαρμόζεται μία μεθοδολογία που αναπτύχθηκε με σκοπό την αξιολόγηση των χαρακτηριστικών των δεδομένων εισόδου ως προς τον σχηματισμό των ομάδων που ορίσθηκαν από τα δύο προηγούμενα βήματα. Στα δύο τελευταία στάδια το νευρο-ασαφές μοντέλο επεξεργάζεται τα δεδομένα ενσωματώνοντας πληροφορίες που προέκυψαν από τα πρώτα στάδια της ανάλυσης. Οι πληροφορίες αυτές αφορούν :

- Τις ομάδες των δεδομένων εισόδου που εντοπίστηκαν. Οι πληροφορίες αυτές χρησιμοποιούνται με στόχο την καλύτερη αρχικοποίηση του μοντέλου

και την προσαρμογή τους στα δεδομένα εισόδου, έτσι ώστε να βελτιωθεί ο χρόνος επεξεργασίας του μοντέλου.

- Τον βαθμό συσχέτισης των χαρακτηριστικών εισόδου με τις ομάδες των δεδομένων εισόδου. Ο βαθμός αυτός χρησιμοποιείται για να διαφοροποιήσει τον ρυθμό μάθησης του νευρο-ασαφούς μοντέλου στους κόμβους-νευρώνες του μοντέλου που αντιστοιχούν στις ομάδες των δεδομένων.

Χρησιμοποιώντας τις πληροφορίες αυτές βελτιώνεται ο χρόνος εκπαίδευσης αλλά και απόδοση του νευρο-ασαφούς μοντέλου. Το υβριδικό αυτό σύστημα είναι ιεραρχικής μορφής αφού η επεξεργασία των δεδομένων γίνεται διαδοχικά από τα διάφορα μέρη του συστήματος, χωρίς ανάδραση πληροφοριών από ένα μέρος του συστήματος προς κάποιο προηγούμενο.

Η αρχιτεκτονική και η λειτουργικότητα του νευρο-ασαφούς μοντέλου βασίζεται στο μοντέλο ARANFIS [61], το οποίο εφαρμόζει την λογική RAN [62] στο μοντέλο SuPFuNIS [60]. Η αρχιτεκτονική περιλαμβάνει τρία επίπεδα: το επίπεδο εισόδου, το επίπεδο των κανόνων και το επίπεδο εξόδου.

Έστω ότι το σύστημα αποτελείται από n εισόδους, p εξόδους και $q(t)$ κρυφούς κόμβους στο χρονικό βήμα t, καθώς ο αριθμός των κόμβων αυτών είναι χρονικά μεταβαλλόμενος αφού το μοντέλο είναι προσαρμοζόμενης δομής. Στο επίπεδο της εισόδου, οι αριθμητικές είσοδοι μετατρέπονται σε ασαφείς μεταβλητές, καθώς οι κόμβοι εισόδου του δικτύου λειτουργούν ως ασαφοποιητές των χαρακτηριστικών των δεδομένων εισόδου. Οι συνδέσεις προς και από τους κόμβους του δικτύου υλοποιούνται από συναρτήσεις συμμετοχής Gauss που ορίζονται από μία τιμή για το κέντρο και μία τιμή διασποράς για τις συναρτήσεις Gauss.

Η διαδικασία ασαφοποίησης των αριθμητικών εισόδων μετατρέπει τις εισόδους σε συναρτήσεις συμμετοχής Gauss με κέντρο x_i^c την τιμή της εισόδου και διασπορά x_i^σ προσαρμοζόμενη από το σύστημα. Εφόσον υφίσταται ήδη γνώση με συμβολική μορφή κανόνων, τότε αυτή μπορεί εύκολα να ενσωματωθεί στο σύστημα με την μορφή κανόνων if-then, οι οποίοι υλοποιούνται με την χρήση κρυφών κόμβων. Όταν όμως εισάγονται νέοι κρυφοί κόμβοι τότε όλες οι συνδέσεις του δικτύου μεταξύ των κόμβων εισόδου και των νέων κόμβων πρέπει να προσδιοριστούν κατά την διάρκεια της εκπαίδευσης. Η τιμή ενεργοποίησης z_j του κρυφού κόμβου j υπολογίζεται ως εξής:

$$z_j = \prod_{i=1}^{n} E_{ij} \tag{4.1}$$

όπου η μεταβλητή E_{ij} αναπαριστά την σχετική επικάλυψη, δηλαδή το ποσοστό του εμβαδού της επιφάνειας τομής των συναρτήσεων συμμετοχής της εισόδου i και της σύνδεσης w_{ij} ως προς το άθροισμα των εμβαδών των δύο επιφανειών [60].

Η απόκριση του κόμβου k στο επίπεδο εξόδου δίνεται από την παρακάτω εξίσωση

$$y_k(t) = \frac{\prod_{j=1}^{q(t)} z_j v_{jk}^c v_{jk}^s}{\prod_{j=1}^{q(t)} z_j v_{jk}^s} \tag{4.2}$$

όπου v_{ij}^c, v_{ij}^σ είναι το κέντρο και η διασπορά αντίστοιχα των συνδέσεων μεταξύ των κόμβων στο επίπεδο των κανόνων και το επίπεδο της εξόδου.

4.2.1 Αξιολόγηση χαρακτηριστικών

Ο σκοπός της μεθοδολογίας ομαδοποίησης με χρήση του μοντέλου των αυτοοργανούμενων χαρτών που περιγράφηκε στο προηγούμενο κεφάλαιο δεν είναι μόνο η αυτόματη ομαδοποίηση των νευρώνων του εκπαιδευμένου χάρτη και κατ' επέκταση και των δεδομένων εισόδου με τα οποία εκπαιδεύτηκε αλλά και αξιολόγηση των χαρακτηριστικών των δεδομένων ως προς τις ομάδες που παράχθηκαν από αυτά. Για να γίνει η αξιολόγηση αυτή η μεθοδολογία αυτή επαναλαμβάνεται τόσες φορές όσες και το πλήθος των χαρακτηριστικών εισόδου, δηλαδή η διάσταση του χώρου των δεδομένων εισόδου. Σε κάθε επανάληψη ένα από τα χαρακτηριστικά εισόδου ευνοείται σε σχέση με τα υπόλοιπα. Αυτό πραγματοποιείται με την μεγέθυνση του χαρακτηριστικού. Δηλαδή οι τιμές των διανυσμάτων των νευρώνων του ευνοημένου χαρακτηριστικού πολλαπλασιάζονται με έναν σταθερό παράγοντα κατά των υπολογισμό της ευκλείδειας απόστασης μεταξύ δύο διανυσμάτων. Επομένως υπολογίζεται μία σταθμισμένη ευκλείδεια απόσταση.

$$\|\mathbf{w}_i - \mathbf{w}_j\| = \sqrt{\sum_n \left((w_{in} - w_{jn}) \cdot \mathbf{a}_n\right)^2}, \quad n = 1, 2, ..., m \tag{4.3}$$

όπου m η διάσταση του χώρου των δεδομένων εισόδου. Το διάνυσμα στάθμισης α είναι διάστασης m και έχει την μορφή : $a = [1, 1, ..., A, ..., 1]$, οπού όταν η διαδικασία ευνοεί το n - οστό χαρακτηριστικό, τότε η τιμή του διανύσματος για αυτήν την διάσταση είναι A με A μεγαλύτερο της μονάδας. Τροποποιώντας με αυτόν τον τρόπο τον υπολογισμό της απόστασης μεταξύ των διανυσμάτων,

είναι επακόλουθο σε κάθε επανάληψη της διαδικασίας με διαφορετικό ευνοημένο χαρακτηριστικό κάθε φορά να προκύπτει και μία διαφορετική ομαδοποίηση των νευρώνων του πλέγματος.

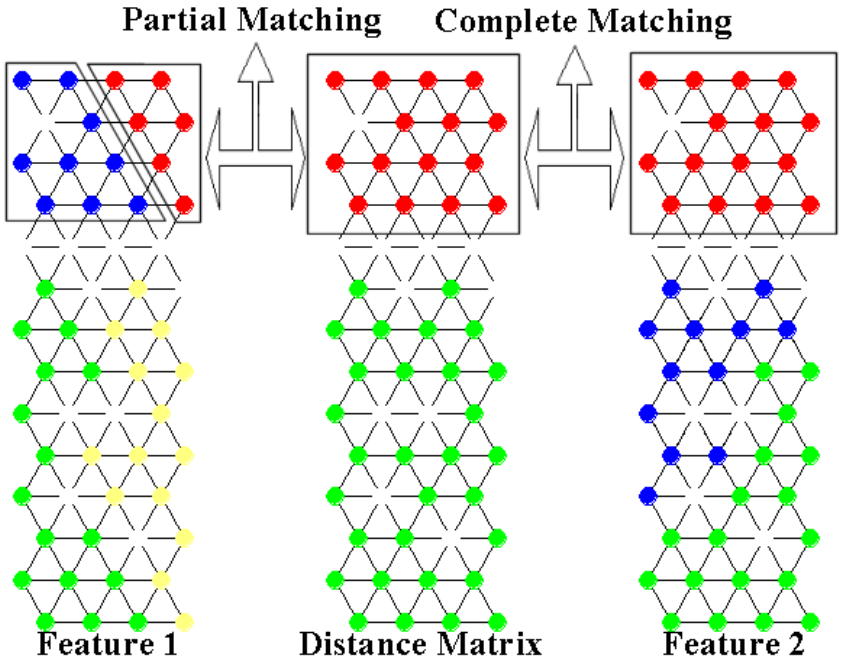
Από την σύγκριση της εκάστοτε νέας ομαδοποίησης με την αρχική ομαδοποίηση που προέκυψε χωρίς την στάθμιση ή ισοδύναμα με διάνυσμα στάθμισης με όλα του τα στοιχεία ίσα με την μονάδα, προκύπτει και αξιολόγηση του εκάστοτε χαρακτηριστικού. Η κάθε ομάδα της εκάστοτε ομαδοποίησης συγκρίνεται με κάθε ομάδα της αρχικής ομαδοποίησης. Από τον αριθμό των κοινών στοιχείων που περιέχουν οι δύο ομάδες προκύπτει και ο βαθμός ομοιότητας. Ο μέγιστος βαθμός ομοιότητας πού προκύπτει από την σύγκριση όλων των ομάδων της εκάστοτε ομαδοποίησης με μία ομάδα της αρχικής ομαδοποίησης αποτελεί και τον βαθμό αξιολόγησης του εκάστοτε χαρακτηριστικού ως προς αυτήν την ομάδα της αρχικής ομαδοποίησης. Σύμφωνα με τα παραπάνω για την ομάδα A της αρχική ομαδοποίησης και το χαρακτηριστικό n, ο βαθμός αξιολόγησης $V_{A,n}$ προκύπτει από την παρακάτω εξίσωση:

$$V_{A,n} = \max_{T} \left(\frac{card(S_A \cap S_{Tj,n})}{card(S_A \cup S_{Tj,n})} \right), \quad j = 1, 2, ..., K \tag{4.4}$$

Στην παραπάνω εξίσωση, T_j είναι μία ομάδα από την ομαδοποίηση που προέκυψε με την στάθμιση υπέρ του χαρακτηριστικού n και K είναι ο αριθμός των ομάδων από την ομαδοποίηση αυτή, ενώ S_A και $S_{Tj,n}$ είναι τα σύνολα που περιέχουν τα διανύσματα των νευρώνων της ομάδας A και T αντίστοιχα. Ο συνάρτηση card επιστρέφει το πλήθος των στοιχείων του συνόλου. Όπως φαίνεται και στο σχήμα 4.1 μία ομάδα από την αρχική ομαδοποίηση μπορεί να συμπίπτει ολικά ή μερικά με κάποια από τις ομάδες που προκύπτουν με την διαφορετική στάθμιση των χαρακτηριστικών. Στην περίπτωση της ολικής ταύτισης, ο βαθμός αξιολόγησης του χαρακτηριστικού ως προς την ομάδα είναι ο μέγιστος δυνατός και όπως προκύπτει και από την εξίσωση (4.4) ίσος με την μονάδα.

4.2.2 Αρχικοποίηση του συστήματος

Η αρχικοποίηση των ελεύθερων παραμέτρων του συστήματος γίνεται με την χρήση των αποτελεσμάτων της μεθοδολογίας ομαδοποίησης με χάρτες SOM. Τα κέντρα των συνδέσεων προς τους κρυφούς κόμβους αρχικοποιούνται με τις τιμές των κέντρων των ομάδων που προέκυψαν από την ομαδοποίηση, ενώ τα κέντρα των συνδέσεων προς τους κόμβους εξόδου παράγονται με χρήση της πληροφορίας της κατηγορίας του κάθε δεδομένου εισόδου και την ανεύρεση των ποσοστών συμμετοχής της κάθε κατηγορίας στις ομάδες. Οι διασπορές αρχικοποιούνται τυχαία σε ένα προκαθορισμένο εύρος τιμών για να διατηρηθεί η



Σχήμα 4.1: *Παράδειγμα μερικής και ολικής ταύτισης μεταξύ της αρχικής ομαδοποίησης και των ομαδοποιήσεων που προέκυψαν από δύο χαρακτηριστικά εισόδου.*

δυνατότητα του συστήματος για καλύτερή γενίκευση.

Η διαδικασία εκπαίδευσης του συστήματος υλοποιείται μέσω της τεχνικής gradient descent. Το κριτήριο του τετραγωνικού σφάλματος χρησιμοποιείται σαν μέτρο επίδοσης της διαδικασίας εκπαίδευσης, η οποία για το χρονικό βήμα t υπολογίζεται ως εξής:

$$e(t) = \frac{1}{2}\sum_{k=1}^{p}(d_k(t) - y_k(t))^2 \tag{4.5}$$

όπου $d_k(t)$ και $y_k(t)$ είναι η επιθυμητή απόκριση του και η πραγματική απόκριση του συστήματος αντίστοιχα για το κόμβο εξόδου k. Το σφάλμα υπολογίζεται για όλες τις εξόδους p και για κάθε πρότυπο εισόδου $x(t)$. Οι ελεύθερες παράμετροι του συστήματος, δηλαδή τα κέντρα και οι διασπορές των συνδέσεων καθώς και οι διασπορές για τα χαρακτηριστικά εισόδου τροποποιούνται με βάση τις εξισώσεις ενημερώσεις, των οποίων η γενική μορφή είναι:

$$u(t+1) = u(t) - \eta(t) \cdot \beta_{ij}\frac{\partial e(t)}{\partial u(t)} \tag{4.6}$$

όπου $\eta(t)$ είναι ο προσαρμοζόμενος ρυθμός μάθησης και είναι μία παράμετρος που εισάγεται από την παρούσα μεθοδολογία. Οι αναλυτικές εξισώσεις των μερικών παραγώγων περιγράφονται στο [61].

4.2.3 Μεταβαλλόμενος ρυθμός μάθησης

Η τιμή του ρυθμού μάθησης $\eta(t)$ επιδρά σε μεγάλο βαθμό στην απόδοση του συστήματος, καθώς είναι στενά συνδεδεμένη με τον ρυθμό σύγκλισης του συστήματος. Παρατηρήθηκε ότι ο κατάλληλος χειρισμός του ρυθμού μάθησης κατά την διάρκεια της εκπαίδευσης οδηγεί σε αισθητά καλύτερα αποτελέσματα και για αυτό ένας μεγάλος αριθμός μεθόδων έχει προταθεί για τον χειρισμό του [51] , [67]. Στο παρόν σύστημα, ο χειρισμός αυτός πραγματοποιείται μέσω της παραμέτρου β_{ij}, η οποία εκφράζει τον βαθμό σημαντικότητας του χαρακτηριστικού εισόδου x_i σε σχέση με την ομάδα j (κόμβος κανόνα). Πιο συγκεκριμένα χρησιμοποιείται ο βαθμός αξιολόγησης V που υπολογίζεται με την μεθοδολογία που περιγράφηκε στην παράγραφο 4.2.1. Η αντίστροφη τιμή του βαθμού αυτού ισούται με την παράμετρο β.

$$\beta_{ij} = \frac{1}{V_{ij}} \tag{4.7}$$

Με τον τρόπο αυτό επιτυγχάνεται η γρήγορη σύγκλιση των καλών συνδέσεων και η καθυστέρηση των υπολοίπων με στόχο την καλύτερη προσαρμογή από ένα μεγαλύτερο εύρος τιμών.

4.2.4 Προσθήκη νέου κανόνα

Τα δεδομένα εισόδου παρουσιάζονται στο σύστημα με την μορφή ζευγών $(x(t), d(t))$ από διανύσματα εισόδου και επιθυμητής εξόδου αντίστοιχα. Εάν ένα νέο διάνυσμα εισόδου δεν ενεργοποιεί σε σημαντικό βαθμό κανένα από τους κόμβους των κανόνων και το σφάλμα είναι αρκετά μεγάλο, τότε ένας νέος κόμβος κανόνας εισάγεται στην δομή του δικτύου. Πιο αναλυτικά, εάν ισχύουν οι παρακάτω συνθήκες

$$\begin{gathered} |d_k(t) - y_k(t)| > \varepsilon = 0.5 \quad \text{(σφάλμα εξόδου)} \\ \max_j \{z_j\} < d = 0.5 \quad \text{(ενεργοποίηση κανόνων)} \end{gathered} \tag{4.8}$$

τότε ένας νέος κόμβος εισάγεται και ο συνολικός αριθμός τους αυξάνεται. Τα κέντρα των συνδέσεων με τους κόμβους εισόδου αρχικοποιούνται με τις τιμές του διανύσματος $x(t)$, καθώς και οι αντίστοιχες διασπορές αρχικοποιούνται ανάλογα με την απόσταση του νέου κόμβου από τον πλησιέστερο υπάρχοντα κόμβο κανόνα. Με αυτόν τον τρόπο τα δεδομένα εισόδου που πληρούσαν τις παραπάνω συνθήκες είναι πιθανότερο να αντιστοιχιστούν στους νέους κόμβους:

$$v^c_{q(t)k} = d_k(t) - y_k(t), \quad k = 1, ..., p \tag{4.9}$$

όπου p είναι ο αριθμός των εξόδων και $q(t)$ ο νέος κόμβος. Οι διασπορές των συνδέσεων $v^\sigma_{q(t)k}$ αρχικοποιούνται τυχαία στο διάστημα $[\min v^\sigma_{jk}, \max v^\sigma_{jk}]$. Η αύξηση της αυστηρότητας στα κριτήρια εισαγωγής νέου κόμβου επιτυγχάνεται με αύξηση της τιμής της παραμέτρου ε και μείωση της τιμής της παραμέτρου d.

4.3 Πειραματική αξιολόγηση

Το πρωτότυπο αυτό υβριδικό σύστημα αξιολογήθηκε με χρήση δεδομένων που χρησιμοποιούνται ευρέως στη σχετική βιβλιογραφία, με σκοπό τη δυνατότητα σύγκρισης των αποτελεσμάτων με αποτελέσματα από αλλά ερευνητικά μοντέλα. Τα πειραματικά αποτελέσματα έδειξαν ότι το πρωτότυπο αυτό υβριδικό μοντέλο παρουσιάζει βελτίωση των αποτελεσμάτων του νευρο-ασαφούς μοντέλου, όταν αυτό χρησιμοποιείται αυτόνομα. Επίσης, τα αποτελέσματα είναι συγκρίσιμα και με διάφορα άλλα γνωστά μοντέλα ταξινόμησης.

Η απόδοση του συστήματος αξιολογήθηκε σε δύο σύνολα δεδομένων, που έχουν ως κύριο χαρακτηριστικό την σχετικά αυξημένη διάσταση του χώρου των δεδομένων εισόδου. Στα πειράματα που πραγματοποιήθηκαν, το 70% του συνόλου χρησιμοποιήθηκε για την εκπαίδευση του συστήματος και το υπόλοιπο 30% για το αξιολόγηση του συστήματος.

Σύνολο δεδομένων Ιονόσφαιρας (Ionosphere Data)

Τα δεδομένα αυτά προέρχονται από ένα σύστημα RADAR στο Goose Bay, Labrador και αποτελείται από 351 εγγραφές από τις οποίες οι 200 συνήθως (στην βιβλιογραφία) χρησιμοποιούνται για εκπαίδευση του συστήματος τα υπόλοιπα 151 για αξιολόγηση. Επίσης έγιναν και πειράματα με χρήση 250 εγγραφών για εκπαίδευση και για αξιολόγηση χρησιμοποιήθηκε το σύνολο των δεδομένων (resubstitution test).

Ο μέσος όρος 10 πειραμάτων με τις ίδιες αρχικές συνθήκες παρουσιάζεται στους παρακάτω πίνακες. Το σύστημα της ομαδοποίησης με χρήση των αυτο-οργανούμενων χαρτών παρήγαγε 10 αρχικές ομάδες και το νευρο-ασαφές σύστημα χρειάστηκε μόνο 10 εποχές εκπαίδευσης για να συγκλίνει στην τελική λύση καθώς στις επόμενες εποχές η μείωση του σφάλματος ήταν αμελητέα. Ο πίνακας 4.2 παρουσιάζονται και συγκριτικά αποτελέσματα με άλλα γνωστά και αποδεκτά συστήματα κατηγοριοποίησης [1].

Πίνακας 4.1: *Ποσοστά (%) ορθής κατηγοριοποίησης του συστήματος για τα δεδομένα της Ιονόσφαιρας.*

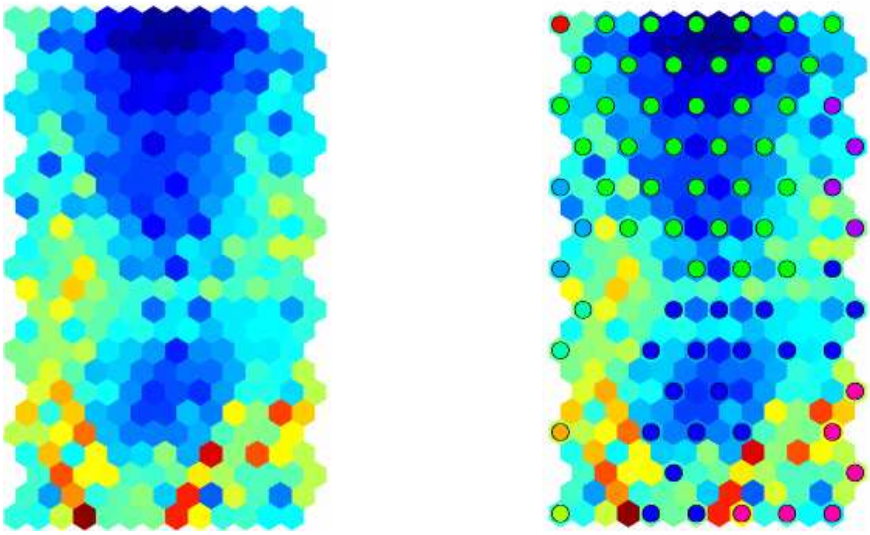
	Τελικός Αριθμός κανόνων		
	13	15	16
201 Εκπαίδευση - 150 Αξιολόγηση	93	94.4	96.67
251 Εκπαίδευση - 351 Αξιολόγηση	92.3	92	92.6

Πίνακας 4.2: *Ποσοστά (%) ορθής κατηγοριοποίησης διαφόρων συστημάτων κατηγοριοποίησης στο σύνολο δεδομένων της Ιονόσφαιρας.*

Σύστημα	Απόδοση(%)
3-NN + simplex	98.7
3-NN	96.7
Our Method	*96.67*
1-NN, Manhattan	96.0
MLP+BP	96.0
C4.5	94.9
SVM	93.2
FSM + rotation	92.8
Linear Perceptron	90.7
CART	88.9

Σύνολο δεδομένων κατάτμησης εικόνων (Image Segmentation)

Το δεύτερο σύνολο δεδομένων αποτελείται από τα χαρακτηριστικά τμημάτων εικόνων. Τα τμήματα ανήκουν σε 7 εικόνες και η τμηματοποίηση είναι χειροποίητη. Κάθε τμήμα αποτελείται από μία περιοχή 3 επί 3 εικονοστοιχείων (pixels) και παράγονται 19 χαρακτηριστικά που περιγράφουν το κάθε τμήμα. Στα πειράματα χρησιμοποιήθηκαν 1610 δεδομένα εισόδου για εκπαίδευση και τα υπόλοιπα 700 για αξιολόγηση. Ο χάρτης SOM ομαδοποιήθηκε σε 4 ομάδες και τα αποτελέσματα του νευρο-ασαφούς συστήματος για 10 και 100 εποχές παρουσιάζονται στο πίνακα 4.3. Και σε αυτή την περίπτωση, τα αποτελέσματα συγκρίνονται με άλλα συστήματα κατηγοριοποίησης στον πίνακα 4.4.



(α) (β)

Σχήμα 4.2: *Χάρτης SOM και ομάδες νευρώνων που προέκυψαν για το σύνολο δεδομένων της Ιονόσφαιρας.*

Πίνακας 4.3: *Ποσοστά (%) ορθής κατηγοριοποίησης του συστήματος για τα δεδομένα των εικόνων.*

Εποχές εκπαίδευση	10	100
Τελικός Αριθμός κανόνων	6	9
Απόδοση εκπαίδευσης (%)	74.22	81.8
Απόδοση αξιολόγησης (%)	72.26	74.91

4.4 Συζήτηση - Συμπεράσματα

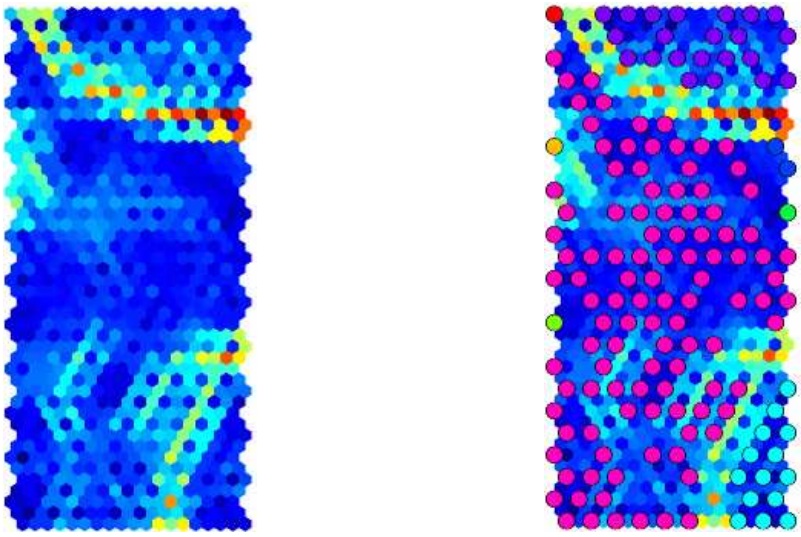
Το πρωτότυπο υβριδικό σύστημα που συγκροτήθηκε από την χρήση της μεθοδολογίας ομαδοποίησης με αυτο-οργανούμενους χάρτες καθώς και την τεχνική αξιολόγησης των χαρακτηριστικών σε συνδυασμό με το νευρο-ασαφές μοντέλο εμφανίζει αποτελέσματα κατηγοριοποίησης βελτιωμένα συγκριτικά με παρεμφερή συστήματα κατηγοριοποίησης.

Ο συνδυασμός αυτός δημιουργεί ένα ιεραρχικό υβριδικό σύστημα. Η ροή της πληροφορίας είναι μονόδρομη και η επεξεργασία των δεδομένων γίνεται σειριακά. Παρόλα αυτά, το υβριδικό αυτό σύστημα αποτελεί επιτυχημένη προσέγγιση καθώς το μοντέλο των αυτο-οργανούμενων χαρτών καταφέρνει σε ικανοποιητικό βαθμό να καλύψει τις αδυναμίες του νευρο-ασαφούς κατηγοριοποιητή και να βελτιώσει την απόδοσή του τόσο σε ποσοστά ορθής κατηγοριοποίησης όσο και σε χρόνους εκπαίδευσης.

Το δεύτερο είναι εξίσου σημαντικό, καθώς οι χρόνοι εκπαίδευσης τέτοιων

Πίνακας 4.4: *Ποσοστά (%) ορθής κατηγοριοποίησης διαφόρων συστημάτων κατηγοριοποίησης στο σύνολο δεδομένων κατάτμησης εικόνων.*

Σύστημα	Απόδοση εκπαίδευσης (%)	Απόδοση αξιολόγησης (%)
Our Method	*95,0*	*90,9*
Itrule	95,6	95,5
Descrim	88,8	88,4
CASTLE	89,2	88,8
RBF	95,3	93,1
Kohonen	95,4	93,3
Backprop	97,2	94,6
C4.5	98,7	96,0



(α) (β)

Σχήμα 4.3: *Χάρτης SOM και ομάδες νευρώνων που προέκυψαν για το σύνολο δεδομένων τμημάτων εικόνων.*

κατηγοριοποιητών παρουσιάζουν μεγάλη εξάρτηση από τις παραμέτρους αρχικοποίησή τους και η κατάλληλη ρύθμισή τους μπορεί να βελτιώσει θεαματικά τους χρόνους αυτούς. Το παραπάνω υβριδικό σύστημα, παρόλο τον πρόσθετο χρόνο επεξεργασίας που εισάγει στην διαδικασία της κατηγοριοποίησης, μειώνει συνολικά τον χρόνο εκπαίδευσης του συστήματος.

Επομένως πληρείτε ένας από τους βασικούς στόχους στον σχεδιασμό ενός υβριδικού συστήματος που είναι η αντιμετώπιση των μειονεκτημάτων του ενός υποσυστήματος από το άλλο.

□

Κεφάλαιο 5

Σύστημα κατηγοριοποίησης βασισμένο στο μοντέλο των κ - πλησιέστερων γειτόνων

5.1 Εισαγωγή

Συνεχίζοντας την έρευνα για την αξιολόγηση της δυνατότητας διαφόρων μοντέλων να ενσωματωθούν σε πρωτότυπα υβριδικά συστήματα, πραγματοποιήθηκε έρευνα στο πεδίο των μοντέλων κατηγοριοποίησης που βασίζονται στην πρόβλεψη μέσω με μίας βάσης γνώσης. Η βασική αρχή αυτής της κατηγορίας αλγορίθμων είναι η απουσία μιας δομής, π.χ. ενός δικτύου νευρώνων, η οποία θα εκπαιδεύεται από τα δεδομένα εισόδου και θα αντιπροσωπεύει τον ευφυή μηχανισμό που περικλείει την γνώση του συστήματος. Τα μοντέλα αυτά χρησιμοποιούν μία βάση γνώσης, δηλαδή δεδομένα τα οποία είναι ήδη κατηγοριοποιημένα και, όταν τους ζητηθεί να αξιολογήσουν ένα πρότυπο άγνωστης κατηγορίας, τότε ανατρέχουν στη βάση γνώσης και προσπαθούν να το ταξινομήσουν χρησιμοποιώντας επιλεγμένα πρότυπα από τη βάση γνώσης. Ο ευφυής μηχανισμός αυτών των συστημάτων εκφράζεται από τη μεθοδολογία που χρησιμοποιείται για την επιλογή των κατάλληλων προτύπων. Σε αυτό το πεδίο, το κυρίαρχο μοντέλο είναι αυτό των «ν - πλησιέστερων γειτόνων». Με βάση αυτό το μοντέλο αναπτύχθηκε ένα πρωτότυπο μοντέλο ταξινόμησης [59],[57],[58].

Μία άλλη γνωστή μεθοδολογία και ευρέως χρησιμοποιούμενη είναι αυτή των naive Bayesian κατηγοριοποιητών [21], [43], [91], η οποία βασίζεται στην θεωρία των πιθανοτήτων. Στην προσέγγιση αυτή υπολογίζεται η posterior πιθανότητα ενός προτύπου να ανήκει σε μία κατηγορία. Ο υπολογισμός αυτός βασίζεται στον αντίστοιχο υπολογισμό εξετάζοντας τα χαρακτηριστικά του προτύπου ξε-

χωριστά. Οι πιθανότητες αυτές συνυπολογίζονται με χρήση της υπόθεσης ότι είναι ανεξάρτητες μεταξύ τους και εξάγεται η ολική πιθανότητα και η κατηγορία του προτύπου. Οι αλγόριθμοι CFP και KNNFP [4], [16], [29], [28] εφαρμόζουν επίσης την ιδέα του ανεξάρτητου υπολογισμού της κατηγορίας ενός νέου προτύπου για κάθε χαρακτηριστικό και στην συνέχεια τον συνδυασμό των επιμέρους αποτελεσμάτων.

5.2 Περιγραφή του μοντέλου

Η μεθοδολογία που αναπτύχθηκε χρησιμοποιεί τέσσερις διαφορετικούς παράγοντες, με στόχο να ταξινομήσει ένα πρότυπο άγνωστης κατηγορίας. Οι παρακάτω παράγοντες αξιολογούνται ξεχωριστά για κάθε σύνολο προτύπων της βάσης γνώσης, στο οποίο ανήκουν όλα τα πρότυπα της ίδιας κατηγορίας.

1. Ο μέγιστος αριθμός των χαρακτηριστικών μεταξύ των προτύπων της βάσης γνώσης, των οποίων οι τιμές - μεμονωμένα - έχουν «μικρή» διαφορά από τις τιμές του αγνώστου προτύπου. Αυτός ο παράγοντας δηλώνει ποιος είναι ο μέγιστος αριθμός χαρακτηριστικών ενός προτύπου της βάσης γνώσης, τα οποία, αν αξιολογηθούν ασυσχέτιστα μεταξύ τους, πλησιάζουν τις αντίστοιχες τιμές των χαρακτηριστικών του άγνωστου προτύπου.

2. Το μέγεθος του συνόλου των προτύπων της βάσης γνώσης, των οποίων ο αριθμός των χαρακτηριστικών που έχει μικρή διαφορά από τις τιμές του αγνώστου προτύπου είναι ίσος με τον μέγιστο αριθμό του προηγούμενου παράγοντα. Με βάση αυτόν τον παράγοντα εντοπίζονται τα πλησιέστερα πρότυπα, λαμβάνοντας υπ' όψιν όχι όλα τα χαρακτηριστικά αλλά όσο το δυνατόν περισσότερα. Με αυτό τον τρόπο, γίνεται προσπάθεια αποφυγής της χρήσης χαρακτηριστικών, τα οποία δεν εμφανίζουν συσχέτιση με την ταξινόμηση του άγνωστου προτύπου

3. Η μέση τιμή των ευκλείδειων αποστάσεων μεταξύ των προτύπων του παραπάνω συνόλου και του άγνωστου προτύπου. Ο παράγοντας αυτός χρησιμοποιείται για να ισοσκελίσει την επίδραση του προηγούμενου παράγοντα.

4. Η μέση τιμή όλων των μικρών διαφορών μεταξύ των τιμών των χαρακτηριστικών των προτύπων της βάσης γνώσης και του άγνωστου προτύπου.

Πιο αναλυτικά αν το διάνυσμα $\mathbf{x}$ αντιστοιχεί σε ένα πρότυπο άγνωστης κατηγορίας και $\mathbf{y}$ είναι ένα πρότυπο που ανήκει στο σύνολο D των προτύπων της

βάσης γνώσης, τότε η ομοιότητα υπολογίζεται με βάση το πλήθος των χαρακτηριστικών που οι διαφορές τους είναι μικρότερες από έναν βαθμό εμπιστοσύνης. Ο αριθμός αυτός είναι το πλήθος των βέβαιων χαρακτηριστικών (Count of Confident Features - $F(\mathbf{x},\mathbf{y})$) μεταξύ των προτύπων $\mathbf{x}$ και $\mathbf{y}$ και υπολογίζεται με την χρήση της συνάρτησης πυρήνα $W(x_i, y_i)$.

$$F(\mathbf{x}, \mathbf{y}) = \sum_i w(x_i, y_i) \tag{5.1}$$

Όπου x_i και y_i είναι οι τιμές του i-στου χαρακτηριστικού των προτύπων $\mathbf{x}$ και $\mathbf{y}$ αντίστοιχα. Η συνάρτηση W ορίζεται ως εξής:

$$W(x_i, y_i) = \begin{cases} 1, & if \quad 1 - \dfrac{|x_i - y_i|}{width_i} \geq Confidence \\ \\ 0, & otherwise \end{cases} \tag{5.2}$$

Μέτρα διασποράς όπως η τυπική απόκλιση των τιμών του χαρακτηριστικού i στο σύνολο D μπορούν να χρησιμοποιηθούν για $width_i$. Η παράμετρος $width$ χρησιμοποιείται για κανονικοποίηση των τιμών των χαρακτηριστικών έτσι ώστε να μπορεί να χρησιμοποιηθεί κοινή τιμή της παραμέτρου $Confidence$ για όλα τα χαρακτηριστικά. Η κανονικοποίηση μπορεί να γίνει και σε ένα στάδιο προεπεξεργασίας των δεδομένων και η παράμετρος $width$ να παραλειφθεί. Η μέγιστη τιμή της παραμέτρου $Confidence$ είναι ίση με την μονάδα και με αυτή την τιμή ο αλγόριθμος επιλέγει μόνο τις τιμές που είναι ακριβώς ίδιες. Αυτή η τιμή μπορεί να χρησιμοποιηθεί σε περίπτωση που κάποια από τα χαρακτηριστικά εισόδου δεν είναι ποσοτικά. Σε αυτή την περίπτωση, εάν υπάρχουν παράλληλα και ποσοτικά χαρακτηριστικά τότε η κανονικοποίηση για τα μη ποσοτικά χαρακτηριστικά παραλείπεται και για τα ποσοτικά εισάγεται διαφορετική τιμή $Confidence$. Η πειραματική αξιολόγηση έδειξε ότι η βέλτιστη απόδοση επιτυγχάνεται για τιμές από 0,2 έως 0,6. Η συνάρτηση $W(x_i, y_i)$ μπορεί να θεωρηθεί και ως μία απλή συνάρτηση πυρήνα, γνωστή και ως Parzen window, που χρησιμοποιείται στον υπολογισμό συναρτήσεων πυκνότητας πιθανότητας με χρήση συναρτήσεων πυρήνα [54].

Η συνθήκη της παραπάνω εξίσωσης μπορεί να επεκταθεί έτσι ώστε να περιοριστεί ακόμα περισσότερο ο αριθμός των χαρακτηριστικών που θεωρούνται παρόμοιες με τιμές των χαρακτηριστικών του πρότυπου $\mathbf{x}$. Ο περιορισμός αυτός επιτυγχάνεται υπολογίζονται τις τιμές της συνάρτησης W για όλα τα πρότυπα της βάσης γνώσης και επιλέγοντας μόνο τα k πλησιέστερα. Η τροποποιημένη συνθήκη είναι:

$$1 - \frac{|x_i - y_i|}{width_i} > Confidence \wedge y_i \in G_k^i(x_i) \tag{5.3}$$

όπου το σύμβολο $\wedge$ αντιστοιχεί στην λογική σύζευξη και το είναι το $G_k^i(x_i)$ σύνολο που περιέχει τις k πλησιέστερες τιμές στην τιμή x_i από το σύνολο D^i. Το σύνολο D^i αποτελείτε από τις τιμές μόνο του χαρακτηριστικού i όλων των προτύπων του συνόλου D (βάση γνώσης).

$$G_k^i(x_i) = \left\{ y_i \in D^i : |x_i - y_i| \leq \min_{y_i \in D^i}{}_k \{|x_i - y_i|\} \right\} \tag{5.4}$$

Ο τελεστής min_k υποδηλώνει την επιλογή των κ-οστού μικρότερου στοιχείου.

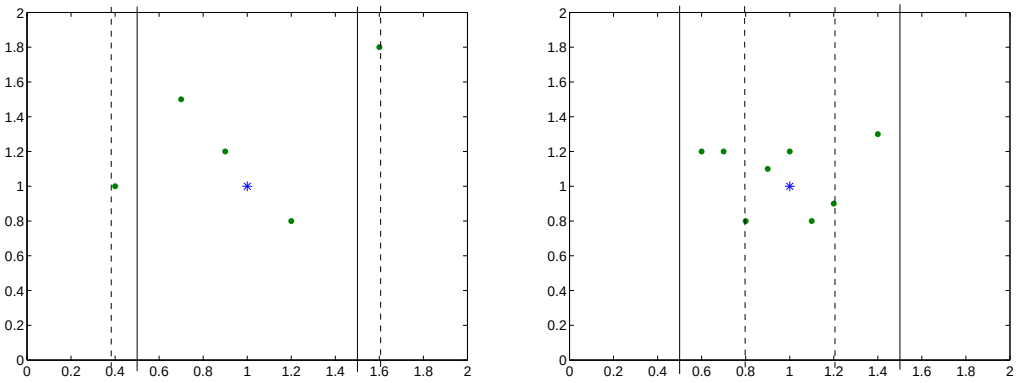
Η χρήση της παραμέτρου k αντιστοιχεί στην τιμή κ του αλγόριθμου των κ πλησιέστερων γειτόνων. Παρόλα αυτά πρέπει να επισημανθεί ότι στην εξίσωση (5.3) και οι δύο συνθήκες εφαρμόζονται σε κάθε χαρακτηριστικό ξεχωριστά. Το γεγονός αυτό αποτελεί σημαντική διαφορά σε σχέση με παρόμοιες μεθοδολογίες, όπως για παράδειγμα ο αλγόριθμος των κ - πλησιέστερων γειτόνων, όπου η συνθήκη των κ - πλησιέστερων γειτόνων εφαρμόζεται με χρήση των αποστάσεων μεταξύ των προτύπων, οι οποίες έχουν υπολογιστεί με βάση τις τιμές όλων των χαρακτηριστικών.

Οι παράμετροι k και $Confidence$ εξυπηρετούν τον ίδιο σκοπό. Θέτουν τα όρια στον αριθμό των τιμών των χαρακτηριστικών που θεωρούνται πλησιέστερες στις τιμές του πρότυπου x, αλλά οι τιμές τους επιδρούν αντιστρόφως ανάλογα σε αυτόν τον σκοπό. Όταν η τιμή του k ελαττώνεται τότε περιορίζονται και τα στιγμιότυπα των χαρακτηριστικών που θεωρούνται γειτονικά ενώ όταν η τιμή του $Confidence$ ελαττώνεται, τα στιγμιότυπα αυξάνονται. Η χρήση της παραμέτρου $Confidence$ επιτρέπει μεταβλητό αριθμό στιγμιότυπων αλλά με μία σταθερή μέγιστη απόσταση ενώ η παράμετρος k επιτρέπει σταθερό αριθμό στιγμιότυπων ανεξαρτήτως απόστασης από την τιμή x_i. Ο τελικός πυρήνας που ορίζεται από την συνάρτηση W αποτελεί την τομή δύο διαφορετικών πυρήνων, ενός πυρήνα σταθερού πλάτους και ενός πυρήνα μεταβλητού πλάτους, το οποίο μεταβάλλεται έτσι ώστε να περιλαμβάνει τις k πλησιέστερες τιμές. Το πλάτος αυτού του σύνθετου πυρήνα είναι:

$$R\left(x_i\right) = \min\left((1 - Confidence) \cdot width_i, \max_{y_i \in G_k^i(x_i)} \left(|x_i - y_i|\right) \right) \tag{5.5}$$

Στο σχήμα 5.1 εμφανίζονται τα όρια όπως ορίζονται από την εξίσωση (5.3).

Στο σχήμα 5.1 τα όρια που ορίζονται από συνθήκη των πλησιέστερων γειτόνων (k=5) διαγράφονται με την διακεκομμένη γραμμή, ενώ αυτά που ορίζονται



Σχήμα 5.1: *Όρια για την οριζόντια διάσταση όπως αυτά ορίζονται από τον βαθμό εμπιστοσύνης και από τον αριθμό των γειτόνων.*

από τον βαθμό εμπιστοσύνης διαγράφονται με συνεχή γραμμή και στις δύο περιπτώσεις τα όρια ορίζονται για το πρώτο χαρακτηριστικό των δεδομένων δηλαδή την οριζόντια διάσταση. Στο σχήμα 5.1 (α), ο βαθμός εμπιστοσύνης περιορίζει τα γειτονικά πρότυπα κατά την οριζόντια διάσταση σε 3, ενώ στο σχήμα 5.1 (β), τα όρια του βαθμού εμπιστοσύνης είναι ευρύτερα από αυτά που ορίζονται μέσω του αριθμού των γειτόνων έτσι ο αριθμός των γειτονικών προτύπων είναι 5. Πρέπει να σημειωθεί ότι και οι δύο συνθήκες εφαρμόζονται σε κάθε χαρακτηριστικό (διάσταση) ξεχωριστά, γεγονός που αποτελεί και την ειδοποιό διαφορά μεταξύ της παρούσας μεθοδολογίας και του αλγόριθμου των κ - πλησιέστερων γειτόνων καθώς στο αλγόριθμο αυτό ο κανόνας των κ - πλησιέστερων γειτόνων εφαρμόζεται αφού υπολογιστούν οι αποστάσεις μεταξύ των προτύπων με χρήση όλων των χαρακτηριστικών τους.

Επομένως η συνάρτηση F που ορίστηκε με χρήση της συνάρτησης $W(x_i, y_i)$ αποτελεί μία μετρική ομοιότητας προσαρμοζόμενη στο δεδομένο περιβάλλον της εκάστοτε βάσης γνώσης. Η μέγιστη τιμή αυτής της μετρικής είναι ο συνολικός αριθμός των χαρακτηριστικών των προτύπων, δηλαδή η διάσταση του χώρου των δεδομένων εισόδου, ενώ η ελάχιστη είναι το μηδέν.

Ακολουθώντας σε αυτό το σημείο την φιλοσοφία του αλγόριθμου των κ - πλησιέστερων γειτόνων πρέπει να υπολογιστεί ένα σύνολο πλησιέστερων γειτόνων του άγνωστου πρότυπου $\mathbf{x}$. Για τον υπολογισμό αυτόν ορίζεται η μέγιστη τιμή του πλήθους των βέβαιων χαρακτηριστικών (maximal Count of Confident Features - $F_i^{max}(\mathbf{x})$) μεταξύ του άγνωστου πρότυπου $\mathbf{x}$ και όλων των προτύπων κάθε κατηγορίας από την βάση γνώσης.

$$F_j^{max}(\mathbf{x}) = \underset{\mathbf{y} \in D_j}{max} \left(F(\mathbf{x}, \mathbf{y}) \right) \tag{5.6}$$

όπου j είναι μία από της κατηγορίες των προτύπων της βάσης γνώσης και D_j είναι το υποσύνολο του συνόλου D που περιέχει όλα τα πρότυπα που ανήκουν στην κατηγορία αυτή. Αυτός ο αριθμός αντιπροσωπεύει για κάθε κατηγορία τον μέγιστο αριθμό των διαφορών των τιμών των χαρακτηριστικών μεταξύ του άγνωστου πρότυπου $\mathbf{x}$ και των πρότυπων της εκάστοτε κατηγορίας που πληρούν τις συνθήκες της εξίσωσης (5.3). Αυτή η μέγιστη τιμή προκύπτει για ένα ή περισσότερα πρότυπα της κάθε κατηγορίας. Χρησιμοποιώντας την τιμή αυτή για την κάθε κατηγορία, ορίζεται ένα σύνολο γειτόνων του άγνωστου πρότυπου για κάθε κατηγορία ως:

$$K_j^{\mathbf{x}} = \left\{ \mathbf{y} \in D_j : F(\mathbf{x}, \mathbf{y}) \geq F_j^*(\mathbf{x}) \right\} \ : \ F_j^*(\mathbf{x}) = F_j^{\max}(\mathbf{x}) - M \tag{5.7}$$

όπου $K_j^{\mathbf{x}}$ είναι το σύνολο των γειτόνων του πρότυπου $\mathbf{x}$ από τα πρότυπα της κατηγορίας j, όπως αυτό ορίζεται από την μεθοδολογία. Εάν υπολογιστεί για όλες τις κατηγορίες της βάσης γνώσης τότε προκύπτει ένα συγκεντρωτικό σύνολο γειτόνων για το άγνωστο πρότυπο. Η μεθοδολογία θεωρεί ότι γείτονες του άγνωστου πρότυπου αποτελούν τα πρότυπα από κάθε κατηγορία, τα οποία επιτυγχάνουν αριθμό βέβαιων χαρακτηριστικών ίσο με τον μέγιστο της κατηγορίας. Η τιμή F_j^* χρησιμοποιείται για να αυξήσει τον αριθμό των προτύπων αυτών που θεωρούνται γειτονικά και θα επηρεάσουν την τελική απόφαση της ταξινόμησης του άγνωστου προτύπου. Αυτό επιτυγχάνεται συμπεριλαμβάνοντας στο σύνολο των γειτόνων και πρότυπα με αριθμό βέβαιων χαρακτηριστικών μικρότερο (τιμή $F(\mathbf{x}, \mathbf{y})$) από τον μέγιστο της κατηγορίας τους (τιμή F_j^{max}). Η παράμετρος M υποδηλώνει την μέγιστη επιτρεπτή διαφορά που μπορεί να έχει ο αριθμός βέβαιων χαρακτηριστικών ενός προτύπου από τον μέγιστο αριθμό της κατηγορίας έτσι ώστε να συμπεριληφθεί στο σύνολο των γειτόνων ($K_j^{\mathbf{x}}$).

Όπως φαίνεται και στην εξίσωση (5.6), η τιμή F_j^{max} υπολογίζεται για κάθε κατηγορία ξεχωριστά και για αυτό δημιουργούνται διαφορετικά σύνολα γειτόνων για κάθε κατηγορία. Με αυτή την προσέγγιση επιτρέπεται να επιλεχθούν ως γείτονες πρότυπα από όλες τις κατηγορίες ακόμα και αν η τιμή F_j^{max} παρουσιάζει μεγάλη διακύμανση για τις διάφορες κατηγορίες. Μια προφανής παραλλαγή της μεθοδολογίας είναι να υπολογιστεί η τιμή F_j^{max} συνολικά για όλα τα πρότυπα της βάσης γνώσης ανεξάρτητα από την κατηγορία στην οποία ανήκουν. Αυτό θα είχε ως αποτέλεσμα την δημιουργία ενός και μοναδικού συνόλου πλησιέστερων γειτόνων όπως και στην προσέγγιση του απλού αλγόριθμου των κ - πλησιέστερων γειτόνων. Ένα βασικό μειονέκτημα της παραλλαγής αυτής είναι ότι υπάρχει η πιθανότητα η τιμή F_j^{max} να υπολογιστεί από ένα πρότυπο μίας κατηγορίας και τα πρότυπα με την αμέσως μικρότερη τιμή $F(\mathbf{x}, \mathbf{y})$ να εμφανίζουν μεγάλη διαφορά στην τιμή αυτή από την τιμή F_j^{max}. Με την χρήση αυτής

της τιμής θα αποκλειστούν πρότυπα από όλες τις άλλες κατηγορίες. Σε αυτή την περίπτωση το σύνολο των πλησιέστερων γειτόνων θα αποτελείται από ελάχιστα πρότυπα και πιθανότητα μόνο ένα, γεγονός που οδηγεί την μεθοδολογία σε λανθασμένες αποφάσεις κατηγοριοποίησης.

5.2.1 Παράγοντες εξισορρόπησης

Ένας επιπλέον παράγοντας που πρέπει να συμπεριληφθεί στην διαδικασία του προσδιορισμού της κατηγορίας ενός άγνωστου προτύπου είναι η μέση απόσταση του άγνωστου πρότυπου από τα πρότυπα του συνόλου των πλησιέστερων γειτόνων για κάθε κατηγορία. Στον υπολογισμό αυτόν χρησιμοποιούνται όλα τα χαρακτηριστικά των προτύπων. Η μέση απόσταση των αυτών των προτύπων (Average Distance of Similar Patterns - $P_j(\mathbf{x})$) για κάθε κατηγορία υπολογίζεται από την παρακάτω εξίσωση.

$$P_j(\mathbf{x}) = \frac{1}{\underset{K_j^{\mathbf{x}}}{mean}\left(\|\mathbf{x}-\mathbf{y}\|\right)} \tag{5.8}$$

Οπού το σύμβολο $\|.\|$ αντιστοιχεί στην ευκλείδεια μετρική ενώ η συνάρτηση $mean()$ υπολογίζει την μέση τιμή. Στην παραπάνω εξίσωση χρησιμοποιείται η αντίστροφη τιμή της μέσης τιμής έτσι ώστε όλοι οι παράγοντες που θα χρησιμοποιηθούν για την απόφαση της κατηγοριοποίησης να έχουν τιμή ανάλογη με την πιθανότητα το άγνωστο πρότυπο να ανήκει σε κάποια κατηγορία.

Σε παραλλαγές του αλγόριθμου των κ - πλησιέστερων γειτόνων, η απόσταση μεταξύ του άγνωστου πρότυπου και ενός προτύπου από το σύνολο των γειτόνων του, έχει χρησιμοποιηθεί για στάθμιση του βάρους του πρότυπου αυτού στην τελική απόφαση κατηγοριοποίησης. Σε αυτές τις περιπτώσεις χρησιμοποιείται η απόσταση του κάθε προτύπου ξεχωριστά. Στη παρούσα μεθοδολογία χρησιμοποιείται η μέση τιμή της απόστασης μεταξύ των προτύπων της κάθε κατηγορίας από το άγνωστο πρότυπο ως παράγοντας στάθμισης όλης της κατηγορίας στην τελική απόφαση. Η επιλογή αυτή έγινε έτσι ώστε αυτή η τιμή να επιδρά εξομαλυντικά και να αντιστοιχεί σε μία συνολική εκτίμηση για όλη την κατηγορία.

Ο δεύτερος παράγοντας εξισορρόπησης που χρησιμοποιεί η παρούσα μεθοδολογία είναι η κανονικοποιημένη διαφορά μεταξύ χαρακτηριστικών του άγνωστου προτύπου και των τιμών των χαρακτηριστικών που πληρούν τις συνθήκες της εξίσωσης (5.3), υπολογιζόμενος για κάθε κατηγορία ξεχωριστά. Δηλαδή εάν η τιμή y_i ικανοποιεί την συνθήκη $W(x_i, y_i) = 1$ τότε η διαφορά της από την τιμή x_i θα συμπεριληφθεί στον υπολογισμό του παράγοντα ασχέτως με το αν το πρότυπο $\mathbf{y}$ συμπεριλαμβάνεται στο σύνολο των πλησιέστερων γειτόνων του

άγνωστου πρότυπου $\mathbf{x}$. Η τιμή του παράγοντα αυτού ορίζεται ως ολική διαφορά των χαρακτηριστικών (All Features Differences - $Q_j(\mathbf{x})$).

$$Q_j(\mathbf{x}) = \left(1 - mean\left(\frac{|x_i - y_i|}{width_i}\right)\right) \cdot \left(1 - std\left(\frac{|x_i - y_i|}{width_i}\right)\right)$$
$$\forall \mathbf{y} \in D_j, \forall i : (W(x_i, y_i) = 1) \tag{5.9}$$

Η συνάρτηση std επιστρέφει την τυπική απόκλιση. Για το εκάστοτε πρότυπο y που συμμετέχει στον υπολογισμό δεν είναι απαραίτητο ότι όλες οι τιμές των χαρακτηριστικών του συμμετέχουν καθώς η συνθήκη $W(x_i, y_i) = 1$ αποτιμάται χωριστά για κάθε χαρακτηριστικό. Στον παράγοντα αυτό, η μέση τιμή των διαφορών συνδυάζεται με την τυπική απόκλισή τους. Ο συνδυασμός γίνεται με τέτοιο τρόπο έτσι ώστε μικρές διαφορές και διαφορές με μεγάλη ομοιομορφία αυξάνουν την τιμή του παράγοντα. Ο σκοπός είναι να επιβραβεύονται οι περιπτώσεις όχι μόνο των μικρών μέσων διαφορών αλλά να αξιολογείται παράλληλα και πόσο ομοιογενείς είναι αυτές. Από την εξίσωση (5.3) προκύπτει ότι η μέγιστη τιμή των κλασμάτων $\frac{|x_i - y_i|}{width_i}$ που θα χρησιμοποιηθούν στην εξίσωση (5.9) προσδιορίζεται από την τιμή που έχει η παράμετρος $Confidence$. Όπως ήδη έχει αναφερθεί, οι τιμές της παραμέτρου αυτής κυμαίνονται μεταξύ 0,2 και 0,6 με τις οποίες το παραπάνω κλάσμα περιορίζεται σε τιμές μικρότερες της μονάδας. Επομένως και οι δύο παράγοντες του γινομένου της εξίσωσης (5.9) θα είναι πάντα θετικοί.

5.2.2 Κανόνας κατηγοριοποίησης

Η τελική απόφαση για την κατηγοριοποίηση ενός άγνωστου πρότυπου λαμβάνεται με συνδυασμό όλων των παραπάνω παραγόντων και τροποποιώντας την γενική εξίσωση κατηγοριοποίησης του αλγόριθμου των κ - πλησιέστερων γειτόνων ως εξής:

$$C_{\mathbf{x}} = \arg\max_j \left\{ Q_j(\mathbf{x}) \cdot P_j(\mathbf{x}) \cdot \sum_{\mathbf{y} \in K_j^{\mathbf{x}}} F(\mathbf{x}, \mathbf{y}) \right\} \tag{5.10}$$

Κάθε παράγοντας που συμμετέχει στην παραπάνω εξίσωση εξυπηρετεί έναν διαφορετικό στόχο. Ο συνδυασμός της ομοιότητας των χαρακτηριστικών, η μέση απόσταση των πλησιέστερων γειτόνων καθώς και το πλήθος των γειτόνων έχει ως σκοπό να προσδιορίσει την κατηγόρια ενός άγνωστου πρότυπου προσεγγίζοντας το πρόβλημα από διαφορετικές πλευρές.

Η μορφή της παραπάνω εξίσωσης παρέχει ακόμα μία αιτιολόγηση για την χρήση μέσων τιμών για τους παράγοντες εξισορρόπησης (εξισώσεις (5.8) και (5.9)). Η μέση τιμή δεν επηρεάζεται άμεσα από το πλήθος των τιμών στις οποίες υπολογίζεται. Το πλήθος των τιμών που συμμετέχουν στους υπολογισμούς των παραπάνω εξισώσεων εξαρτάται από το πλήθος των προτύπων που θεωρούνται γειτονικά στο άγνωστο πρότυπο. Το πλήθος όμως αυτό, το οποίο είναι πολύ σημαντικό στην τελική απόφαση της κατηγοριοποίησης λαμβάνεται υπ' όψιν στην άθροιση των τιμών της συνάρτησης $F(\mathbf{x}, \mathbf{y})$ στην εξίσωση (5.10), επομένως θα ήταν πλεονασμός να επιδρά και στους δύο παράγοντες εξισορρόπησης.

5.2.3 Πολυπλοκότητα

Η πολυπλοκότητα του αλγόριθμο των κ - πλησιέστερων γειτόνων είναι ανάλογη του μεγέθους της βάσης γνώσης, δηλαδή του αριθμού των προτύπων που αυτή περιέχει. Πιο αναλυτικά, ο υπολογισμός των αποστάσεων μεταξύ ενός άγνωστου προτύπου και όλων των προτύπων της βάσης γνώσης απαιτεί n επαναλήψεις όπου στην κάθε μία επανάληψη απαιτούνται d πράξεις, άρα συνολικά $n \cdot d$ πράξεις. Όπου n είναι ο αριθμός των προτύπων στην βάση γνώσης και d είναι ο αριθμός των χαρακτηριστικών εισόδου. Εφόσον έχουν υπολογιστεί οι αποστάσεις, η επιλογή των κ - πλησιέστερων προτύπων είναι ένα πρόβλημα που μπορεί να επιλυθεί σε γραμμικό χρόνο. Επομένως τα συνολικά βήματα εκτέλεσης που απαιτούνται για την κατηγοριοποίησης ενός άγνωστου προτύπου είναι $n \cdot d + n$ και η υπολογιστική πολυπλοκότητα είναι $O(n \cdot d)$ και εάν ισχύει ότι $n \gg d$ τότε η πολυπλοκότητα ελαττώνεται σε $O(n)$.

Στην παρούσα μεθοδολογία απαιτείται ο προσδιορισμός των βέβαιων χαρακτηριστικών, ο οποίος πρέπει να γίνει ανεξάρτητα για κάθε χαρακτηριστικό εισόδου. Αυτό σημαίνει ότι για κάθε χαρακτηριστικό απαιτούνται n βήματα υπολογισμού των διαφορών και την εύρεση των κ - πλησιέστερων διότι όπως αναφέρθηκε και προηγουμένως το πρόβλημα αυτό απαιτεί γραμμικό χρόνο επίλυσης. Επομένως χρειάζονται n βήματα για κάθε χαρακτηριστικό άρα $n \cdot d$ βήματα για όλα τα χαρακτηριστικά. Επιπλέον χρειάζονται ακόμα $n \cdot d$ βήματα για τον υπολογισμό του πλήθους των βέβαιων χαρακτηριστικών για κάθε πρότυπο. Ο υπολογισμός των πλησιέστερων γειτόνων γίνεται ξεχωριστά για κάθε κατηγορία με χρήση του $F_j^{max}(\mathbf{x})$. Για να γίνει αυτό χρειάζεται να εξεταστούν όλα τα πρότυπα της βάσης γνώσης δύο φορές, άρα $2 \cdot n$ βήματα. Μία φόρα για τον προσδιορισμό της τιμής αυτής για κάθε κατηγορία και μία για τον σχηματισμό των συνόλων των πλησιέστερων γειτόνων. Συνολικά απαιτούνται $2 \cdot n \cdot d + 2 \cdot n$ βήματα και η συνολική πολυπλοκότητα είναι επίσης $O(n \cdot d)$. Ο παρακάτω πίνακας

περιέχει συγκεντρωτικά τις φάσεις των αλγορίθμων καθώς και την αντίστοιχη πολυπλοκότητα.

Πίνακας 5.1: *Υπολογιστική πολυπλοκότητα της παρούσας μεθοδολογίας.*

Αλγόριθμος κ - πλησιέστερων γειτόνων		Μεθοδολογία	
Αποστάσεις προτύπων	$O(n \cdot d)$	Αποστάσεις χαρακτηριστικών	$O(n \cdot d)$
Προσδιορισμός γειτόνων	$O(n)$	Προσδιορισμός βέβαιων χαρακτηριστικών	$O(n \cdot d)$
		Προσδιορισμός γειτόνων	$O(n)$
Σύνολο	$O(n \cdot d)$	Σύνολο	$O(n \cdot d)$

Το πρόβλημα της επιλογή των κ - μικρότερων στοιχείων ενός συνόλου, το οποίο εμφανίζεται και εδώ κατά την επιλογή των κ - πλησιέστερων γειτόνων ή των κ - πλησιέστερων χαρακτηριστικών, είναι ένα πρόβλημα του οποίου η επίλυση μπορεί να επιταχυνθεί και να επιτευχθεί σε υπογραμμικούς χρόνους με την χρήση κατάλληλων δομών δεδομένων. Αυτές οι προσεγγίσεις μπορούν να εφαρμοστούν στον αλγόριθμο τον κ - πλησιέστερων γειτόνων αλλά παράλληλα μπορούν να εφαρμοστούν και στη παρούσα μεθοδολογία για τον προσδιορισμό τον κ - πλησιέστερων χαρακτηριστικών οπότε η βελτίωση θα ήταν ανάλογη και στις δύο περιπτώσεις.

5.3 Διαδικασίες προεπεξεργασίας

Για την περαιτέρω βελτίωση της απόδοσης της μεθοδολογίας αναπτύχθηκαν δύο διαφορετικές διαδικασίες προεπεξεργασίας των προτύπων της βάσης γνώσης. Η πρώτη έχει στόχο την αξιολόγηση των προτύπων της βάσης γνώσης και στοχεύει στην βελτίωση της απόδοσης κατηγοριοποίησης και την αντιμετώπιση της παρουσίας προτύπων θορύβου. Η δεύτερη διαδικασία παρέχει ένα δυναμικό σχήμα αξιολόγησης που συνδυάζει την βασική ιδέα της μεθοδολογίας με την χρήση αυτο-οργανούμενων χαρτών.

5.3.1 Αξιολόγηση προτύπων βάσης γνώσης

Σε συνδυασμό με την παραπάνω μεθοδολογία κατηγοριοποίησης, εκτελείται και μια μέθοδος αξιολόγησης των προτύπων της βάσης γνώσης, με στόχο τον αποκλεισμό από την βάση γνώσης των προτύπων outliers. Η μεθοδολογία της αξιολόγησης της βάσης γνώσης ανήκει στο χώρο των μεθόδων ελάττωσης που συχνά

χρησιμοποιούνται σε συνδυασμό με τον αλγόριθμο των κ - πλησιέστερων γειτόνων ή παραλλαγές του.

Για να αξιολογηθεί η βάση γνώσης, για κάθε πρότυπο αυτής υπολογίζεται ένας βαθμός αξιολόγησης. Ο βαθμός αυτός αρχικά είναι μηδέν. Στην βάση γνώσης εφαρμόζεται η μεθοδολογία κατηγοριοποίησης με την τεχνική της εξαίρεσης του ενός (leave-one-out test). Στην τεχνική αυτή ο αλγόριθμος κατηγοριοποίησης εφαρμόζεται τόσες φορές όσες και τα πρότυπα της βάσης γνώσης. Σε κάθε εκτέλεση ένα πρότυπο της βάσης γνώσης εξαιρείται από αυτήν και θεωρείται άγνωστης κατηγορίας και στην συνέχεια κατηγοριοποιείται από τον αλγόριθμο. Έστω για παράδειγμα ότι το τρέχον άγνωστο πρότυπο κατηγοριοποιείται στην κατηγορία j. Εάν ο αλγόριθμος επιτύχει σωστή κατηγοριοποίηση τότε τα πρότυπα της βάσης γνώσης τα οποία ανήκουν στην ίδια κατηγορία και λήφθηκαν υπ' όψιν για το αποτέλεσμα, δηλαδή τα πρότυπα που ανήκουν στο σύνολο $K_j^{\mathbf{x}}$, αυξάνουν τον βαθμό αξιολόγησης τους κατά μία μονάδα. Στην αντίθετη περίπτωση, δηλαδή εάν το πρότυπο δεν κατηγοριοποιηθεί σωστά τότε τα πρότυπα αυτά μειώνουν κατά μία μονάδα τον βαθμό τους. Μετά το τέλος όλης της διαδικασίας, τα πρότυπα με αρνητικό βαθμό αξιολόγησης εξαιρούνται από την βάση γνώσης.

Παρόμοιες μεθοδολογίες με τον ίδιο στόχο όπως η E-NN και All k-NN [75], [84] λειτουργούν με αντίθετο τρόπο, εξαιρούν από την βάση γνώσης τα πρότυπα που δεν κατηγοριοποιούνται σωστά. Η παρούσα μεθοδολογία επιχειρεί να αφαιρέσει από την βάση γνώσης τα πρότυπα τα οποία προκάλεσαν περισσότερες λανθασμένες κατηγοριοποιήσεις. Η παραπάνω διαδικασία εκτελείτε μία φόρα και όχι επαναληπτικά θέτοντας ως στόχο μία τιμή βελτιστοποίησης όπως παρόμοιες μεθοδολογίες.

5.3.2 Ενσωμάτωση των Αυτο-οργανούμενων Χαρτών

Η παρούσα μεθοδολογία βασίζεται στην χρήση του αριθμού των βέβαιων χαρακτηριστικών για κάθε πρότυπο ως μέτρο ομοιότητας. Αυτό που δεν λαμβάνεται υπ' όψιν είναι το ποία είναι κάθε φορά αυτά τα χαρακτηριστικά, τα οποία χρησιμοποιούνται για τον υπολογισμό της ομοιότητας. Με σκοπό να αξιοποιηθεί αυτή η πληροφορία, χρησιμοποιήθηκε το μοντέλο των αυτο-οργανούμενων χαρτών στην υλοποίηση μίας διαδικασίας αξιολόγησης των συνδυασμών των βέβαιων χαρακτηριστικών. Η διαδικασία αυτή σχηματίζει ένα σύνολο δεδομένων, το οποίο αποτελείτε από τους συνδυασμούς των χαρακτηριστικών, κατά την διάρκεια της εκτέλεσης της διαδικασίας αξιολόγησης που περιγράφηκε στην προηγούμενη παράγραφο. Με αυτό το σύνολο εκπαιδεύεται ένας αυτο-οργανούμενος χάρτης, ο

οποίος χρησιμοποιείται στην συνέχεια σαν αξιολογητής των συνδυασμών των χαρακτηριστικών που προκύπτουν.

Πιο αναλυτικά, οι συνδυασμοί προκύπτουν από τον ορισμό της συνάρτησης πυρήνα W στις εξισώσεις (5.2) και (5.3). Έστω ότι για τα πρότυπα $\mathbf{x}$ και $\mathbf{y}$ ορίζεται ένα διάνυσμα $\mathbf{w}$ ως εξής:

$$\mathbf{w_{xy}} = (w_1, w_2, ..., w_d) \tag{5.11}$$

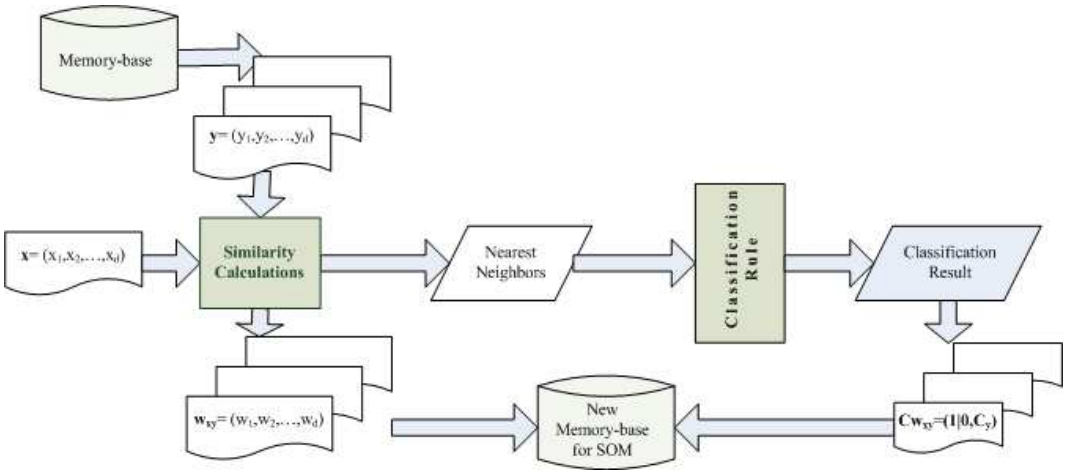
όπου $w_i = W(x_i, y_i)$ και d είναι ο αριθμός των χαρακτηριστικών. Το παραπάνω διάνυσμα υποδηλώνει τον συνδυασμό των χαρακτηριστικών του πρότυπου $\mathbf{x}$ που θεωρούνται «βέβαια» σε σχέση με τα χαρακτηριστικά του πρότυπου $\mathbf{y}$. Ένα διάνυσμα $\mathbf{w_{xy}}$ συνδυασμού των χαρακτηριστικών υπολογίζεται για κάθε πρότυπο που ανήκει στο σύνολο $K_j^{\mathbf{x}}$ που ορίζεται στην εξίσωση (5.7) και θεωρείται το σύνολο των πλησιέστερων γειτόνων, όπως αυτό ορίζεται από την μεθοδολογία.

Ένα ζευγάρι τιμών χρησιμοποιείται για να χαρακτηρίσει κάθε διάνυσμα $\mathbf{w_{xy}}$. Η πρώτη τιμή z αντιστοιχεί στην ορθότητα της τελικής απόφασης κατηγοριοποίησης του πρότυπου $\mathbf{x}$ και οι δυνατές τιμές είναι 1 ή 0, για σωστή και λανθασμένη κατηγοριοποίηση αντίστοιχα. Η δεύτερη τιμή είναι η κατηγορία $C_{\mathbf{y}}$ του πρότυπου $\mathbf{y}$ της βάσης γνώσης, το οποίο μετά την σύγκριση του με το άγνωστο πρότυπο $\mathbf{x}$ σχημάτισε το διάνυσμα $\mathbf{w_{xy}}$.

$$C_{\mathbf{w_{xy}}} = \langle z, C_{\mathbf{y}} \rangle \tag{5.12}$$

Μετά το τέλος της διαδικασίας αξιολόγησης όλων των προτύπων, τα διανύσματα $\mathbf{w}$, δηλαδή οι συνδυασμοί των χαρακτηριστικών, συγκροτούν ένα νέο σύνολο δεδομένων, το οποίο θα χρησιμοποιηθεί για την εκπαίδευση ενός αυτο-οργανούμενου χάρτη. Ο στόχος είναι μέσω του αυτο-οργανούμενου χάρτη να βρεθούν ομάδες όμοιων διανυσμάτων $\mathbf{w}$ άρα κατ' επέκταση και όμοιων συνδυασμών χαρακτηριστικών. Κατά την εκτέλεση της διαδικασίας αξιολόγησης, η σύγκριση κάθε φορά μεταξύ των άγνωστων προτύπων και των προτύπων της βάσης γνώσης οδηγεί στην δημιουργία μίας μεγάλης ποικιλίας διαφορετικών συνδυασμών χαρακτηριστικών, ιδίως όταν το σύνολο δεδομένων είναι μεγάλης διάστασης εισόδου. Ο αυτο-οργανούμενος χάρτης παρέχει την δυνατότητα κβαντισμού των συνδυασμών σε ένα μικρότερο αριθμό. Μετά την εκπαίδευση του χάρτη, χρησιμοποιείται και πληροφορία του χαρακτηρισμού των διανυσμάτων $\mathbf{w}$ (εξ. 5.12). Η πληροφορία αυτή θα χαρακτηρίσει κάθε νευρώνα του εκπαιδευμένου χάρτη.

Μετά την ολοκλήρωση της εκπαίδευσης του αυτο-οργανούμενου χάρτη, κάθε συνδυασμός χαρακτηριστικών (διάνυσμα $\mathbf{w}$) αντιστοιχίζεται σε έναν νευρώνα



Σχήμα 5.2: *Μπλοκ διάγραμμα της διαδικασίας αξιολόγησης.*

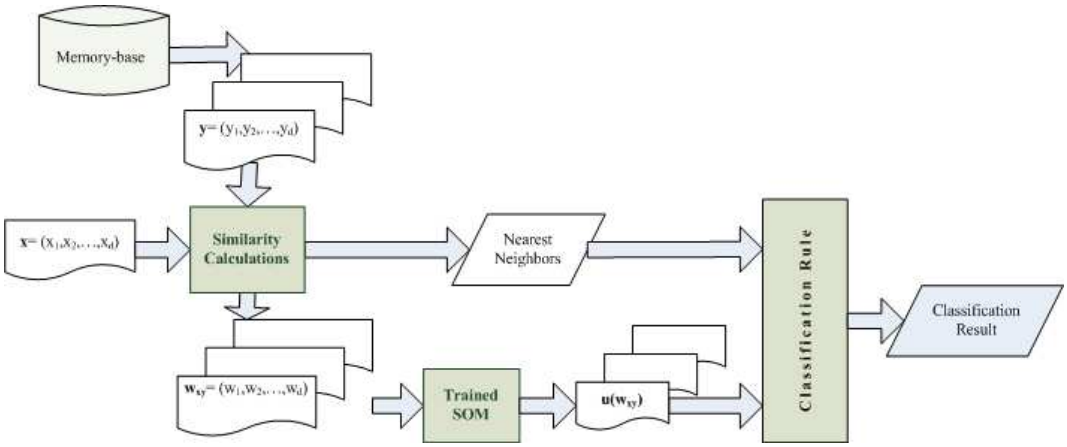
νικητή (BMU). Ο κάθε νευρώνας μπορεί να χαρακτηριστεί από αυτούς τους συνδυασμούς. Δηλαδή, ο κάθε νευρώνας συγκεντρώνει ένα πλήθος από συνδυασμούς που χρησιμοποιήθηκαν για σωστή ή λάθος κατηγοριοποίηση όπως προκύπτει από την τιμή z του ζεύγους τιμών που χαρακτηρίζει τον κάθε συνδυασμό (εξ. 5.12). Ένας παράγοντας αξιολόγησης u για τον i-οστό νευρώνα σε σχέση με την κλάση j υπολογίζεται διαιρώντας τον αριθμό των σωστών κατηγοριοποιήσεων με τον συνολικό αριθμό των κατηγοριοποιήσεων της συγκεκριμένης κατηγορίας:

$$u_i^j = \frac{|\{\mathbf{w} \in U_i, C_\mathbf{w} = \langle 1, j \rangle\}|}{|\{\mathbf{w} \in U_i, C_\mathbf{w} = \langle z, j \rangle\}|} \tag{5.13}$$

όπου U_i είναι το σύνολο των διανυσμάτων $\mathbf{w}$ που έχουν αντιστοιχιστεί στον νευρώνα i του εκπαιδευμένου χάρτη και ο τελεστής $|\cdot|$ υποδηλώνει το πλήθος ενός συνόλου.

Το σχήμα 5.2 απεικονίζει ένα μπλοκ διάγραμμα της διαδικασίας αξιολόγησης που χρησιμοποιείται για την δημιουργία του συνόλου των συνδυασμών που εκπαιδεύσουν το μοντέλο του αυτο-οργανούμενου χάρτη.

Μετά το τέλος αυτής της διαδικασίας, ο εκπαιδευμένος χάρτης μπορεί να λειτουργήσει ως αξιολογητής συνδυασμών που θα προκύψουν κατά την διαδικασία ταξινόμησης νέων άγνωστων προτύπων. Ο παράγοντας αξιολόγησης που έχει υπολογιστεί για τον κάθε νευρώνα μπορεί να χρησιμοποιηθεί για την αξιολόγηση του συνδυασμού αυτού σε περίπτωση που αυτός χρησιμοποιηθεί στην κατηγοριοποίηση του άγνωστου προτύπου. Αυτή η αξιολόγηση βασίζεται στην προηγούμενη εμπειρία της χρήσης παρόμοιων συνδυασμών, η οποία αντιπροσωπεύεται από τους νευρώνες του εκπαιδευμένου χάρτη και τους παράγοντες αξιολόγησης που τους αντιστοιχούν. Όπως έχει περιγραφεί και παραπάνω, η



Σχήμα 5.3: *Μπλοκ διάγραμμα της μεθόδου με την χρήση των αυτο-οργανούμενων χαρτών για αξιολόγηση των συνδυασμών των χαρακτηριστικών.*

παρούσα μεθοδολογία υπολογίζει τους πλησιέστερους γείτονες ενός άγνωστου προτύπου έτσι ώστε να είναι δυνατή η κατηγοριοποίηση του. Με τον ίδιο τρόπο όπως και κατά την διάρκεια της διαδικασίας αξιολόγησης, προκύπτουν συνδυασμοί χαρακτηριστικών. Αυτοί οι συνδυασμοί μπορούν να χαρακτηριστούν από την κατηγορία του προτύπου που κάθε φορά συγκρίθηκε με το άγνωστο πρότυπο. Στην συνέχεια αντιστοιχίζεται ο κάθε συνδυασμός με το νευρώνα νικητή στον εκπαιδευμένο χάρτη. Η διαφορά είναι ότι η ορθότητα του αποτελέσματος της κατηγοριοποίησης είναι άγνωστη και πρέπει να εκτιμηθεί.

Χρησιμοποιώντας τον εκπαιδευμένο χάρτη σε κάθε συνδυασμό χαρακτηριστικών που αντιπροσωπεύεται από ένα διάνυσμα $\mathbf{w}$ ανατίθεται ο βαθμό αξιολόγησης u του νευρώνα στον οποίο ανήκει και της αντίστοιχης κατηγορίας.

$$u(\mathbf{w_{xy}}) = u_i^j, \text{if} \mathbf{w_{xy}} \in U_i \wedge C_\mathbf{y} = j \tag{5.14}$$

Η τιμή $u(\mathbf{w_{xy}})$ μπορεί να χρησιμοποιηθεί σαν βάρος που θα επηρεάσει την επίδραση του πρότυπου $\mathbf{y}$ στο τελικό αποτέλεσμα της κατηγοριοποίησης. Επομένως κανόνας κατηγοριοποίησης (5.10) μπορεί να τροποποιηθεί ως εξής:

$$C_\mathbf{x} = \arg\max_j \left\{ Q_j(\mathbf{x}) \cdot P_j(\mathbf{x}) \cdot \sum_{\mathbf{y} \in K_j^\mathbf{x}} u(\mathbf{w_{xy}}) \cdot F(\mathbf{x}, \mathbf{y}) \right\} \tag{5.15}$$

Με την χρήση του παράγοντα αξιολόγησης με αυτόν τον τρόπο παρέχεται ένα σύστημα δυναμικής αξιολόγησης των πιθανών συνδυασμών που προκύπτουν κατά την διάρκεια της διαδικασίας κατηγοριοποίησης. Το σχήμα 5.3 απεικονίζει ένα μπλοκ διάγραμμα της διαδικασίας κατηγοριοποίησης σε συνδυασμό με την μέθοδο αξιολόγησης με χρήση του αυτο-οργανούμενου χάρτη.

5.4 Πειραματική αξιολόγηση

Με στόχο την αξιολόγηση της απόδοσης της μεθοδολογίας σε προβλήματα κατηγοριοποίησης πραγματοποιήθηκε πειραματική μελέτη με χρήση συνόλων δεδομένων (benchmarks) που χρησιμοποιούνται ευρέως στο ερευνητικό πεδίο. Τα σύνολα αυτά δεδομένων περιγράφουν προβλήματα από διαφορετικούς χώρους και ποικίλουν ως προς το πλήθος των προτύπων, τον αριθμό των χαρακτηριστικών και των κατηγοριών (Παράρτημα Α).

Όλα τα πειράματα πραγματοποιήθηκαν με την μέθοδο 10-fold cross-validation και τα αποτελέσματα τα οποία καταγράφονται παρακάτω είναι η μέση τιμή 10 εκτελέσεων για κάθε σύνολο δεδομένων. Παρατίθενται αποτελέσματα της μεθοδολογίας σε συνδυασμό με την μέθοδο αξιολόγησης με την χρήση των αυτο-οργανούμενων χαρτών αλλά και χωρίς αυτήν. Τα αποτελέσματα συγκρίνονται με αποτελέσματα από τον αλγόριθμο των κ - πλησιέστερων γειτόνων με χρήση της Ευκλείδειας απόστασης αλλά και με παραλλαγές του με χρήση διαφορετικών μετρικών απόστασης (HOEM, HVDM, DVDM, IVDM, WVDM). Επίσης περιλαμβάνονται και αποτελέσματα από τους αλγόριθμους ENN και ALLk-NN, οι οποίοι αποτελούν δύο αναγνωρισμένες μεθοδολογίες ελάττωσης. Τα αποτελέσματα από τις παραπάνω μεθοδολογίες πηγάζουν από τις εργασίες [85],[87]. Σε αυτές τις εργασίες δεν περιλαμβάνονται η τυπική απόκλιση των αποτελεσμάτων αλλά τα πειράματα εκτελέστηκαν ακολουθώντας την ίδια στρατηγική, επομένως είναι άμεσα συγκρίσιμα με τα αποτελέσματα από την παρούσα πειραματική μελέτη. Τα αποτελέσματα από την εκτέλεση των πειραμάτων περιέχονται στον πίνακα 5.2.

Στον πίνακα αυτόν έχουν σημειωθεί οι υψηλότερες αποδόσεις για κάθε σύνολο δεδομένων. Σε επτά από τα δέκα σύνολα δεδομένων η παρούσα μεθοδολογία χωρίς την χρήση των αυτο-οργανούμενων χαρτών εμφανίζει καλύτερα αποτελέσματα από όλες τις υπόλοιπες μεθόδους. Η χρήση της μεθόδου αξιολόγησης των συνδυασμών των χαρακτηριστικών με το μοντέλο των αυτο-οργανούμενων χαρτών βελτιώνει ακόμα περισσότερο τα αποτελέσματα της μεθοδολογίας. Συνολικά η παρούσα μεθοδολογία με ή χωρίς τους αυτο-οργανούμενους χάρτες επιτυγχάνει καλύτερα αποτελέσματα σε οκτώ σύνολα δεδομένων.

Ο πίνακας 5.3 περιέχει τις μέσες τιμές της μέγιστης τιμής του πλήθους των βέβαιων χαρακτηριστικών $F_i^{max}(\mathbf{x})$ (εξ. 5.6) για κάθε σύνολο δεδομένων. Οι τιμές αυτές είναι ενδεικτικές του γεγονότος ότι η μεθοδολογία κατ την διαδικασία της κατηγοριοποίησης δεν χρησιμοποιεί απαραίτητα όλα τα χαρακτηριστικά, αλλά μόνο αυτά που θεωρούνται γειτονικά σε κάθε περίπτωση.

Στην πειραματική αξιολόγηση περιλήφθηκαν 4 ακόμα σύνολα δεδομένων τα

Πίνακας 5.2: *Ποσοστά ορθής κατηγοριοποίησης*

Σύνολα δεδομένων	Απόδοση (%)($\sqrt{s^2}$)							
	k - NN, Eucl.	k - NN, HOEM	k - NN, HVDM	k - NN, DVDM	k - NN, IVDM	k - NN, WVDM	ENN	All-KNN	Παρούσα μέθοδος	Παρούσα μέθοδος (SOM)
Breast Cancer	94.99	95.28	94.99	95.57	95.57	95.57	97.00	97.00	97.02 (1.12)	**97.28** (1.07)
Glass	**72.36**	70.52	**72.36**	56.06	70.54	71.49	65.91	67.75	71.40 (3.57)	69.40 (4.03)
Image Segment.	92.86	93.57	92.86	92.38	92.86	93.33	91.90	92.14	**95.58** (1.22)	94.63 (1.20)
Ionosphere	86.32	86.33	86.32	92.60	91.17	91.44	84.04	84.05	89.90 (3.85)	**94.02** (3.14)
Iris	94.67	95.33	94.67	92.00	94.67	96.00	95.33	95.33	**96.67** (2.05)	**96.67** (2.12)
Liver Disorders	62.92	63.47	62.92	55.04	58.23	57.09	61.12	60.24	68.54 (4.52)	**69.01** (4.12)
Pima Indians	71.09	70.31	71.09	71.89	69.28	70.32	**75.39**	74.88	73.92 (5.32)	74.75 (5.18)
Sonar	87.02	86.60	87.02	78.45	84.17	84.19	81.79	80.36	**90.93** (3.72)	89.07 (3.56)
Vehicle	70.93	70.22	70.93	63.72	69.27	65.37	69.52	70.21	74.01 (4.92)	**75.12** (4.15)
Wine	95.46	95.46	95.46	94.38	97.78	97.22	94.93	94.93	97.83 (1.02)	**98.24** (1.56)

Πίνακας 5.3: *Μέση τιμή* $F_i^{max}(\boldsymbol{x})$

Σύνολα δεδομένων	$F_i^{max}(\mathbf{x})$	Αριθμός χαρακτηριστικών
Breast Cancer	8.20	9
Glass	8.14	9
Image Segment.	14.55	19
Ionosphere	28.09	34
Iris	3.98	4
Liver Disorders	5.27	6
Pima Indians	6.82	8
Sonar	49.99	60
Vehicle	16.74	18
Wine	9.27	13

οποία χαρακτηρίζονται είτε από την μεγάλη διάσταση εισόδου είτε από μεγάλο αριθμό πρότυπων. Η χρήση αυτών των συνόλων έγινε με σκοπό να εξεταστεί η δυνατότητα κλιμάκωσης της παρούσας μεθοδολογίας σε δεδομένα υψηλών διαστάσεων. Τα δύο πρώτα σύνολα ("Internet Advertisements" & "Covertype") είναι από το UCI machine-learning repository [8] ενώ τα δύο επόμενα ("Arcene" & "Gisette") προέρχονται από το NIPS Feature Selection Challenge [27]. Τα πειράματα διενεργήθηκαν και σε αυτή την περίπτωση με την μέθοδο 10-fold cross-validation και η σύγκριση των αποτελεσμάτων γίνεται με αποτελέσματα από την εκτέλεση του απλού αλγόριθμου των κ - πλησιέστερων γειτόνων στα ίδια σύνολα δεδομένων ακολουθώντας την ίδια στρατηγική εκτέλεσης. Για την επιλογή της παραμέτρου κ στην εφαρμογή του αλγόριθμου των κ - πλησιέστερων γειτόνων, πραγματοποιήθηκαν δοκιμές με διάφορες τιμές και επιλέχθηκε αυτή που κάθε φορά παρήγαγε τα καλύτερα αποτελέσματα. Τα αποτελέσματα περιγράφονται στον πίνακα 5.4 μαζί με μία σύντομη περιγραφή των χαρακτηριστικών των συνόλων των δεδομένων.

Όπως φαίνεται και από τον πίνακα 5.4, στα δύο πρώτα σύνολα δεδομένων η μεθοδολογία σε συνδυασμό με την χρήση των αυτο-οργανούμενων χαρτών παρουσίασε καλύτερα αποτελέσματα σε σύγκριση με τον αλγόριθμο των κ - πλησιέστερων γειτόνων, ενώ στα επόμενα δύο τα αποτελέσματα της παρούσας μεθοδολογίας ήταν καλύτερα από αυτά των κ - πλησιέστερων γειτόνων ακόμα και χωρίς την χρήση του μοντέλου των αυτο-οργανούμενων χαρτών.

Για την περαιτέρω τεκμηρίωση της πειραματικής αξιολόγησης της μεθοδολο-

Πίνακας 5.4: *Ποσοστά ορθής κατηγοριοποίησης συνόλων υψηλών διαστάσεων*

			Απόδοση (%)($\sqrt{s^2}$)		
Σύνολα δεδομένων	Αριθμός προτύπων	Αριθμός χαρακτ.	k - NN, Eucl.	Παρούσα μέθοδος	Παρούσα μέθοδος (SOM)
Internet Adv.	2359	1558	95.68 (1.37)	94.36 (1.22)	**96.12** (1.89)
Covertype	20000	54	83.14 (1.03)	83.05 (1.24)	**85.22** (1.52)
Arcene	200	10000	85.50 (4.51)	86.97 (4.73)	**89.92** (4.21)
Gisette	1000	5000	91.31 (2.59)	93.04 (2.48)	**95.57** (2.46)

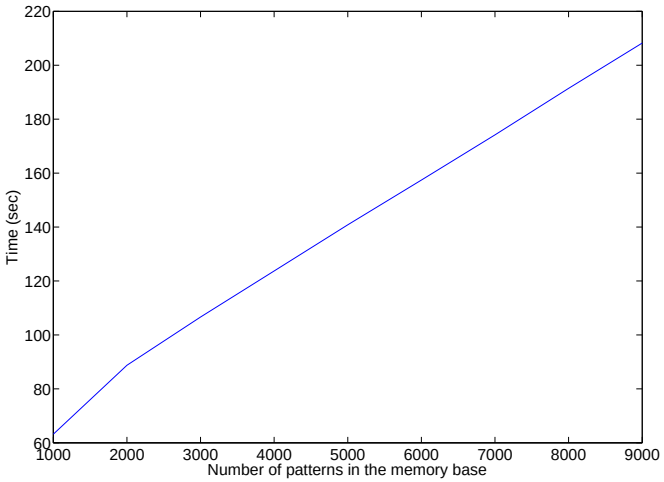
γίας, υπολογίστηκαν τα διαστήματα εμπιστοσύνης των αποτελεσμάτων. Στον πίνακα 5.5 καταγράφονται τα διαστήματα εμπιστοσύνης των διαφορών μεταξύ των αποτελεσμάτων της παρούσας μεθοδολογίας με και χωρίς την χρήση των αυτο-οργανούμενων χαρτών και των αποτελεσμάτων του αλγόριθμου των κ - πλησιέστερων γειτόνων με χρήση της Ευκλείδειας απόστασης. Σε όλες τις περιπτώσεις η διαφορά μεταξύ των αποτελεσμάτων εμφανίζεται στατιστικά σημαντική, ακόμα και στις περιπτώσεις που η παρούσα μεθοδολογία εμφανίζει αποτελέσματα μικρότερα από αυτά του αλγόριθμου των κ - πλησιέστερων γειτόνων (αρνητικές μέσες τιμές).

Στην παράγραφο 5.2.3 παρουσιάστηκε μία σύντομη ανάλυση της πολυπλοκότητας της μεθόδου. Συμπληρωματικά σε αυτήν την ανάλυση παρατίθενται παρακάτω δύο διαγράμματα που αφορούν τους χρόνους εκτέλεσης της μεθοδολογίας. Στο διάγραμμα 5.4 απεικονίζονται ο χρόνος εκτέλεσης της μεθοδολογίας για την ταξινόμηση 1000 προτύπων του συνόλου δεδομένων "Covertype". Οι χρόνοι απεικονίζονται σαν συνάρτηση του μεγέθους της βάσης γνώσης, η οποία είχε μεταβλητό αριθμό προτύπων ξεκινώντας από 1000 πρότυπα και με μέγιστη τιμή τα 9000 πρότυπα. Όπως φαίνεται και από το διάγραμμα οι χρόνοι εκτέλεσης αυξάνονται γραμμικά σε σχέση με τον μέγεθος της βάσης γνώσης, γεγονός που επιβεβαιώνει την ανάλυση της παραγράφου 5.2.3.

Στο διάγραμμα 5.5 απεικονίζεται η σύγκριση των χρόνων εκτέλεσης της μεθοδολογίας χωρίς την μέθοδο αξιολόγησης των χαρακτηριστικών με χρήση των αυτο-οργανούμενων χαρτών. Απεικονίζονται χρόνοι για όλα τα σύνολα δεδομένων που χρησιμοποιήθηκαν στην παραπάνω πειραματική ανάλυση. Η επάνω ράβδος για κάθε σύνολο δεδομένων αντιστοιχεί στον λόγο του χρόνου εκτέλεσης της μεθοδολογίας ως προς τον χρόνο εκτέλεσης του αλγόριθμου των κ - πλησιέστερων γειτόνων. Η δεύτερη ράβδος αντιστοιχεί στον χρόνο εκτέλεσης του αλγόριθμου των κ - πλησιέστερων γειτόνων ως προς τον εαυτό του οπότε και η τιμή της είναι πάντα ίση με την μονάδα και απεικονίζεται μόνο για λόγους σύγκρισης. Οι χρόνοι εκτέλεσης της μεθοδολογίας εμφανίζονται από

Πίνακας 5.5: *Διαστήματα εμπιστοσύνης*

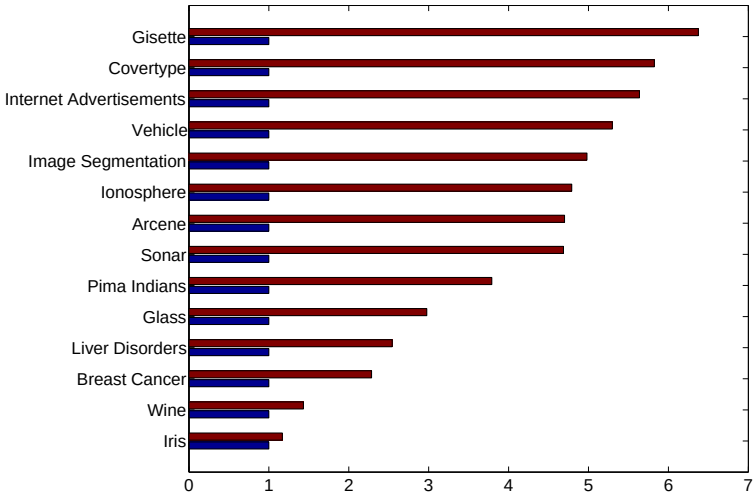
	Διαφορά μέσων τιμών(±)	
Σύνολα δεδομένων	Παρούσα μέθοδος	Παρούσα μέθοδος (SOM)
Breast Cancer	2,03 (0,05)	2,29 (0,05)
Glass	-0,96 (0,14)	-2,96 (0,14)
Image Segmentation	2,72 (0,05)	1,77 (0,05)
Ionosphere	3,58 (0,12)	7,70 (0,11)
Iris	2,00 (0,05)	2,00 (0,05)
Liver Disorders	5,62 (0,13)	6,09 (0,13)
Pima Indians	2,83 (0,15)	3,66 (0,15)
Sonar	3,91 (0,13)	2,05 (0,12)
Vehicle	3,08 (0,13)	4,19 (0,12)
Wine	2,37 (0,04)	2,78 (0,04)
Internet Advertisements	-1,32 (0,04)	0,44 (0,05)
Covertype	-0,09 (0,03)	2,08 (0,04)
Arcene	1,47 (0,13)	4,42 (0,12)
Gisette	1,73 (0,07)	4,26 (0,07)



Σχήμα 5.4: *Χρόνος εκτέλεσης σε σχέση με το μέγεθος της βάσης γνώσης για το σύνολο "Covertype"*

σχεδόν ίσοι έως περίπου 6 φορές μεγαλύτεροι από τους χρόνους εκτέλεσης του αλγόριθμου των κ - πλησιέστερων γειτόνων. Τα πειραματικά αποτελέσματα που

απεικονίζονται στα διαγράμματα επιβεβαιώνουν την ανάλυση της παραγράφου 5.2.3.



Σχήμα 5.5: *Σύγκριση χρόνων παρούσας μεθοδολογίας και αλγόριθμου κ - πλησιέστερων γειτόνων*

5.4.1 Ρύθμιση Παραμέτρων

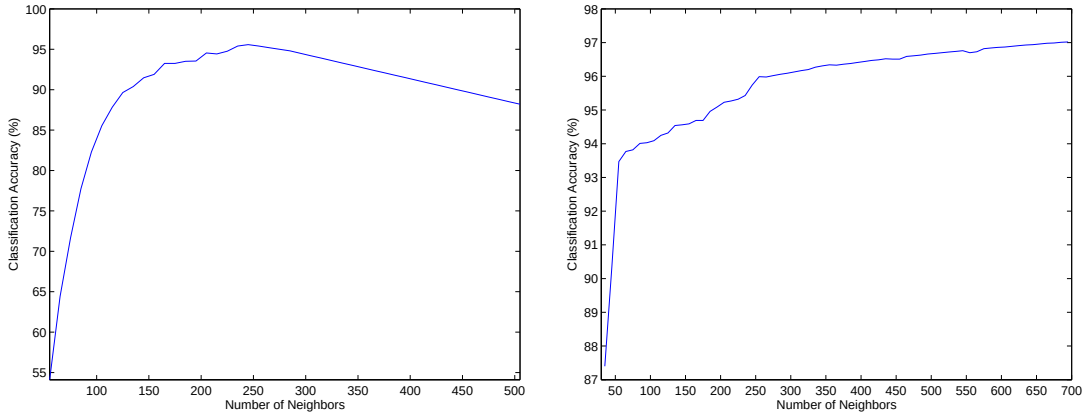
Η παράμετρος "Confidence" είχε σε όλα τα πειράματα την σταθερή τιμή 0.5 καθώς και παράμετρος M (εξ. 5.7) είχε την σταθερή τιμή 2. Τα αποτελέσματα μπορούν να βελτιωθούν ακόμα περισσότερο εάν γίνει μία εξαντλητική αναζήτηση συνδυασμού τιμών στον χώρο των παραμέτρων αυτών. Παρόλα αυτά επιλέχθηκε να μην γίνει αυτό έτσι ώστε να αναδειχθεί το γεγονός ότι και χωρίς αυτήν την εξαντλητική αναζήτηση και παραμετροποίηση της μεθόδου, τα αποτελέσματα είναι εξίσου ικανοποιητικά σε σύγκριση με τις υπόλοιπες μεθοδολογίες. Η μόνη παράμετρος που είχε μεταβλητή τιμή για τα διάφορα τιμή είναι η παράμετρος k και οι τιμές αυτές καταγράφονται στον πίνακα 5.6. Οπού στον πίνακα αυτόν η τιμή καταγράφεται ως «Όλες» υποδηλώνει ότι μόνο η παράμετρος "Confidence" χρησιμοποιήθηκε για τον προσδιορισμό του συνόλου των βέβαιων χαρακτηριστικών. Οι τιμές της παραμέτρου αυτής είναι υψηλότερες από εκείνες της αντίστοιχης τιμής του αλγόριθμου των κ - πλησιέστερων γειτόνων που συνήθως επιτυγχάνει καλύτερα αποτελέσματα για τιμές μικρότερες του 10.

Τα επόμενα δύο γραφήματα απεικονίζουν την απόδοση της μεθοδολογίας σε σχέση με την τιμή της παραμέτρου k ενδεικτικά για δύο από τα σύνολα δεδομένων.Όπως φαίνεται από τα γραφήματα, η απόδοση παρουσιάζει μία αύξουσα

Πίνακας 5.6: *Τιμές παραμέτρου* k

Σύνολα δεδομένων	Παράμετρος k
Breast Cancer	Όλες
Glass	Όλες
Image Segment.	245
Ionosphere	165
Iris	40
Liver Disorders	50
Pima Indians	Όλες
Sonar	100
Vehicle	215
Wine	50

πορεία έως μία μέγιστη τιμή και στην συνέχεια φθίνει ή παρουσιάζει μία σχετικά μονότονα αυξητική πορεία έως το ανώτατο όριο της τιμής της παραμέτρου k που είναι ο αριθμός των προτύπων στην βάση γνώσης.Και στις δύο περιπτώσεις η εξέλιξη των τιμών είναι σχετικά ομαλή και δεν παρουσιάζει έντονες διακυμάνσεις και τοπικά ελάχιστα ή μέγιστα. Το γεγονός αυτό υποδηλώνει ότι η ρύθμιση της παραμέτρου για την επίτευξη της μεγίστης απόδοσης είναι σχετικά απλή διαδικασία.



(α) (β)

Σχήμα 5.6: *Απόδοση σε σχέση με την τιμή της παραμέτρου* k. *(α) Σύνολο δεδομένων image segmentation. (β) Σύνολο δεδομένων breast cancer.*

Για την χρήση των αυτο-οργανούμενων χαρτών χρησιμοποιήθηκε η βιβλιοθήκη λογισμικού SOM Toolbox [79]. Η εκπαίδευση του χάρτη απαιτεί την ρύθ-

μιση κάποιων παραμέτρων όπως το μέγεθος του χάρτη και ο αριθμός των εποχών εκπαίδευσης. Η συγκεκριμένη βιβλιοθήκη παρέχει ευριστικές διαδικασίες για τον αυτόματο καθορισμό των παραμέτρων αυτών που αναλύονται στην αναφορά [79]. Σύμφωνα με αυτήν την αναφορά, ο αριθμός S των νευρώνων του χάρτη καθορίζεται από την παρακάτω εξίσωση.

$$S = 5\sqrt{n} \tag{5.16}$$

Όπου n είναι ο αριθμός των προτύπων εισόδου που θα χρησιμοποιηθούν για την εκπαίδευση του χάρτη. Η αναλογία μεταξύ των διαστάσεων του χάρτη s_1 και s_2 καθορίζεται ως:

$$\frac{s_1}{s_2} = \sqrt{\frac{e_1}{e_2}} \tag{5.17}$$

όπου e_1 και e_2 ($e_1 < e_2$) είναι οι δύο μεγαλύτερες ιδιοτιμές του πίνακα των δεδομένων εισόδου. Καθώς η χρήση του χάρτη αποτελεί βοηθητικό επίπεδο στην όλη διαδικασία, οι παραπάνω ευριστικές μέθοδοι υιοθετήθηκαν με σκοπό την απλοποίηση της διαδικασίας και την αποφυγή περισσότερων παραμέτρων ρυθμιζόμενων από τον χρήστη.

5.5 Συζήτηση - Συμπεράσματα

Στο κεφάλαιο αυτό παρουσιάστηκε μία πρωτότυπη μεθοδολογία κατηγοριοποίησης, η οποία σε συνδυασμό με την μεθοδολογία αξιολόγησης του συνδυασμού των χαρακτηριστικών αποτελεί ένα νέο υβριδικό σύστημα. Το υβριδικό αυτό σύστημα είναι ένα σύστημα από αλληλεπιδρώντα μοντέλα καθώς η μέθοδος κατηγοριοποίησης τροφοδοτεί με δεδομένα τον αυτο-οργανούμενο χάρτη κατά την διαδικασία της αξιολόγησης, ενώ κατά την διαδικασία της κατηγοριοποίησης ενός άγνωστου πρότυπου ζητείται αξιολόγηση για τους νέους συνδυασμούς των χαρακτηριστικών, η οποία επιστρέφεται από τον αυτο-οργανούμενο χάρτη για να αξιοποιηθεί στον κανόνα κατηγοριοποίησης. Επομένως η ροή της πληροφορίας μεταξύ των δύο συστημάτων είναι αμφίδρομη.

Τα αποτελέσματα από την πειραματική μελέτη υποδεικνύουν ότι η παρούσα μεθοδολογία επιτυγχάνει ικανοποιητικά αποτελέσματα κατηγοριοποίησης. Ο απλός αλγόριθμος των κ - πλησιέστερων γειτόνων προσπαθεί να επιλύσει το πρόβλημα της κατηγοριοποίησης υπολογίζοντας τοπικά όρια κατηγοριοποίησης για κάθε άγνωστο πρότυπο. Η γενική ιδέα της παρούσας μεθοδολογίας είναι να εφαρμοστεί η παραπάνω προσέγγιση σε κάθε χαρακτηριστικό ξεχωριστά. Με

βάση αυτόν τοπικό ορισμό της ομοιότητας μεταξύ των τιμών των χαρακτηριστικών, ορίζεται ένα μέτρο ομοιότητας μεταξύ των προτύπων που προκύπτει από τον αριθμό των χαρακτηριστικών που πληρούν τα κριτήρια ομοιότητας. Η προσέγγιση αυτή μπορεί να θεωρηθεί και ως μία υλοποίηση της ανθρώπινης προσέγγισης στο πρόβλημα της σύγκρισης αντικειμένων.

Αποτελεί συνηθισμένη πρακτική, όταν συγκρίνονται δύο αντικείμενα, να μετρώνται τα χαρακτηριστικά των αντικειμένων αυτών που εμφανίζουν ομοιότητα. Εάν το πρόβλημα αντιμετωπισθεί στην βάση της σύγκρισης ενός αντικειμένου σε σχέση με τα αντικείμενα μίας συλλογής, είτε αυτή είναι παρούσα είτε είναι καταχωρημένη στην μνήμη ενός ανθρώπου, τότε τα κριτήρια ομοιότητας που τίθενται για την σύγκριση μεταξύ των αντικειμένων είναι σχετικά με τις τιμές εκάστοτε χαρακτηριστικού στο σύνολο των αντικειμένων. Αυτός ο ανθρώπινος τρόπος προσέγγισης είναι αρκετά όμοιος με τον τρόπο που η παρούσα μεθοδολογία προσεγγίζει το πρόβλημα. Βέβαια αυτός ο τρόπος εκτίμησης αγνοεί την κατηγορία των προτύπων αλλά ούτως ή άλλως αυτός είναι βασικό χαρακτηριστικό των μεθοδολογιών που βασίζονται σε μία βάση γνώσης.

Παρόλο που η παρούσα μεθοδολογία δεν αποτελεί ξεκάθαρα μία μέθοδο στάθμισης ή επιλογής χαρακτηριστικών, μπορεί να εξεταστεί κάτω από αυτό το πρίσμα, καθώς υπολογίζει τα πλησιέστερα πρότυπα πραγματοποιώντας μία επιλογή χαρακτηριστικών. Η επιλογή ενός συγκεκριμένου χαρακτηριστικού γίνεται με βάση την ομοιότητα που εμφανίζουν οι τιμές του χαρακτηριστικού αυτού από τα πρότυπα μεταξύ των οποίων πραγματοποιείται η σύγκριση. Με βάση την κατηγοριοποίηση των μεθόδων που παρουσιάστηκε στο παράγραφο 2.3.1, αυτού του είδους η επιλογή χαρακτηριστικών μπορεί να χαρακτηριστεί ως μία ενσωματωμένη μέθοδος φίλτρο. Ο χαρακτηρισμός της ως φίλτρο βασίζεται στο γεγονός ότι η επιλογή γίνεται πριν τον τελικό καθορισμό των πλησιέστερων προτύπων και δεν έχει καμία ανατροφοδότηση από την τελική έκβαση της κατηγοριοποίησης. Επίσης αυτός ο τρόπος επιλογή των χαρακτηριστικών αποτελεί μία "univariate" προσέγγιση καθώς τα κριτήρια και η απόφαση για την επιλογή του κάθε χαρακτηριστικού λειτουργούν ανεξάρτητα από τα υπόλοιπα χαρακτηριστικά.

Με στόχο να ξεπεραστεί το μειονέκτημα αυτό, δύο διαδικασίες έχουν ενσωματωθεί στην μεθοδολογία. Η πρώτη είναι ο υπολογισμός της μέσης απόστασης των πλησιέστερων προτύπων προς το άγνωστο πρότυπο (Average Distance of Similar Patterns - $P_j(\mathbf{x})$) για κάθε κατηγορία. Στον υπολογισμό αυτής της απόστασης λαμβάνουν μέρος όλες οι τιμές των χαρακτηριστικών των προτύπων. Όπως επισημάνθηκε στην παράγραφο 5.2.2, η χρήση αυτής της τιμής ομαλοποιεί την απόδοση του αλγόριθμου.

Η δεύτερη διαδικασία είναι η χρήση των αυτο-οργανούμενων χαρτών για την

αξιολόγηση των συνδυασμών των χαρακτηριστικών. Η διαδικασία αυτή μπορεί να χαρακτηριστεί ως «πολυμεταβλητή» (multivariate) διαδικασία περιτυλίξεως (wrapper). Διότι σε αυτήν την διαδικασία αξιολογούνται συνδυασμοί χαρακτηριστικών και όχι μεμονωμένα χαρακτηριστικά. Επίσης η διαδικασία πρέπει να λάβει ανατροφοδότηση από την μέθοδο κατηγοριοποίησης κατά το στάδιο της προεπεξεργασίας έτσι ώστε να μπορούν να χαρακτηριστούν οι νευρώνες του εκπαιδευμένου χάρτη με βάση το αποτέλεσμα της κατηγοριοποίησης και την κατηγορία των προτύπων. Ο εκπαιδευμένος χάρτης μπορεί να αξιολογήσει πιθανούς συνδυασμούς χαρακτηριστικών που προκύπτουν κατά την σύγκριση ενός νέου άγνωστου προτύπου και των προτύπων της βάσης γνώσης. Αυτό το σχήμα στάθμισης είναι βασισμένο στην ιδέα της δημιουργίας ενός νέου συνόλου από συνδυασμούς χαρακτηριστικών που έχουν ήδη χρησιμοποιηθεί στο παρελθόν με γνωστή την ορθότητα του αποτελέσματος της κατηγοριοποίησης. Ο αυτο-οργανούμενος χάρτης χρησιμοποιείται στην συνέχεια ως εκτιμητής πιθανότητας.

Παρόλο που αυτή η διαδικασία αξιολόγησης εισάγει ένα πρόσθετο υπολογιστικό κόστος στην συνολική μεθοδολογία, αυτό το κόστος προστίθεται κυρίως στην διάρκεια της προεπεξεργασίας. Επίσης, σε σύγκριση με άλλες μεθοδολογίες περιτυλίξεως, οι οποίες ακολουθούν μία επαναληπτική διαδικασία επιλογής χαρακτηριστικών, αξιολόγησης και επαναεπιλογής, η διαδικασία της δημιουργίας του αυτο-οργανούμενου χάρτη εκτελείται μία φόρα και το επιπρόσθετο υπολογιστικό κόστος μπορεί να υπολογιστεί εκ των προτέρων με βάση την υπολογιστική πολυπλοκότητα της εκπαίδευσης του αυτο-οργανούμενου χάρτη [40].

□

Κεφάλαιο 6

Αυτο-οργανούμενοι χάρτες και συμβολική αναπαράσταση γνώσης

Στο κεφάλαιο αυτό παρουσιάζονται δύο μεθοδολογίες, οι οποίες εκμεταλλεύονται τα πλεονεκτήματα που παρουσιάζει το μοντέλο των αυτο-οργανούμενων χαρτών με σκοπό να μετασχηματίσουν την γνώση που ενσωματώνεται σε έναν εκπαιδευμένο χάρτη σε συμβολική μορφή. Στην πρώτη μεθοδολογία [10], ο μετασχηματισμός αυτός αφορά την δημιουργία συμβολικών καταστάσεων που χρησιμοποιούνται για την δημιουργία πιθανοτικών μοντέλων, ενώ στην δεύτερη περίπτωση παρουσιάζεται μία προσέγγιση σε μεθόδους εξαγωγής κανόνων από έναν εκπαιδευμένο χάρτη [55].

6.1 Υβριδικό σύστημα αναγνώρισης χειρονομιών

6.1.1 Εισαγωγή

Εάν τα δεδομένα εισόδου σχηματίζουν χρονικές ακολουθίες τότε το μοντέλο των αυτο-οργανούμενων χαρτών αδυνατεί να ενσωματώσει αυτήν την πληροφορία στην δομή του. Για να αντιμετωπιστεί αυτό το πρόβλημα έχουν αναπτυχθεί παραλλαγές του αρχικού μοντέλου. Το μοντέλο RSOM (Recurrent SOM) [44] αποτελεί μία αναδρομική μορφή του αρχικού μοντέλου που ακολουθεί τις βασικές αρχές των αναδρομικών νευρωνικών δικτύων, ενώ στην εργασία [68] οι συγγραφείς παρουσιάζουν ένα μοντέλο αυτο-οργανούμενου χάρτη όπου κάθε νευρώνας αντιπροσωπεύεται από μια σειρά διανυσμάτων και με αυτόν τον τρόπο είναι δυνατή η επεξεργασία δεδομένων αυτής της μορφής καθώς το μοντέλο έχει επίγνωση της διαδοχής των δεδομένων εισόδου. Το μοντέλο RSOM χρησιμοποιήθηκε σε συνδυασμό με μοντέλα Markov στην αναγνώριση της κίνησης

φοιτητών μέσα στον χώρο ενός πανεπιστημίου [33].

Όμως μία από τις εργασίες στις οποίες μπορούν να χρησιμοποιηθούν οι αυτο-οργανούμενοι χάρτες είναι στον διανυσματικό κβαντισμό ενός συνόλου δεδομένων. Μετά την εκπαίδευση του μοντέλου, οι νευρώνες του χάρτη λειτουργούν ως στάθμες κβαντισμού του χώρου εισόδου. Η δυνατότητα αυτή μπορεί αξιοποιηθεί σε περιπτώσεις όπου απαιτείται ο μετασχηματισμός ενός συνεχούς χώρου εισόδου σε ένα διακριτό χώρο έτσι ώστε να είναι δυνατή, για παράδειγμα, η εφαρμογή πιθανοτικών μοντέλων όπως τα μοντέλα Markov. Σε αυτή την περίπτωση, οι νευρώνες ενός εκπαιδευμένου αυτο-οργανούμενου χάρτη μπορούν να χρησιμοποιηθούν και ως σύμβολα-καταστάσεις τα οποία θα αποτελέσουν την βάση για την δημιουργία ενός πιθανοτικού μοντέλου καταστάσεων. Τέτοια μοντέλα χρησιμοποιούνται σε προβλήματα όπου τα δεδομένα εισόδου χαρακτηρίζονται από χρονική διαδοχή. Μοντέλα όπως οι αλυσίδες Markov που χρησιμοποιούνται κατά κόρον σε αυτές τις περιπτώσεις, απαιτούν τα δεδομένα εισόδου να έχουν την μορφή διακριτών συμβολικών καταστάσεων.

Στην συνέχεια παρουσιάζεται ένα υβριδικό σύστημα αναγνώρισης χειρονομιών που υλοποιεί την παραπάνω θεώρηση. Η αναγνώριση χειρονομιών είναι ένα πεδίο αυξανόμενης ερευνητικής δραστηριότητας στο πλαίσιο της ευρύτερης έρευνας στην επικοινωνία-διαδραστικότητα ανθρώπου μηχανής. Η αναγνώριση χειρονομιών είτε αφορά την κίνηση του χεριού ή/και την χειρομορφή είναι μία διαδικασία που στο σύνολο της εμπλέκει τεχνικές και μεθοδολογίες από τα πεδία της επεξεργασίας πολυμεσικής πληροφορίας, την επεξεργασία σήματος, την μηχανική μάθηση, στατιστική ανάλυση, κ.α.

Αναλυτική επισκόπηση διαφόρων μεθοδολογιών παρουσιάζεται στα [52] και [89].Στο πρώτο, οι συγγραφείς παρουσιάζουν κυρίως θέματα αναγνώρισης νοηματικής γλώσσας και επικεντρώνονται κυρίως στο θέμα του προσδιορισμού της θέσης του χεριού. Στην δεύτερη εργασία, οι συγγραφείς αναλύουν το πρόβλημα της περιγραφής της χειρομορφής και της εξαγωγής χαρακτηριστικών από σύνολα εικόνων.

Οι μεθοδολογίες κατηγοριοποίησης περιλαμβάνουν διάφορες τεχνικές ανάλογα με τα εξαγόμενα χαρακτηριστικά και το στάδιο της επεξεργασίας στο οποίο χρησιμοποιούνται. Τέτοιες τεχνικές είναι τα μοντέλα HMM (Hidden Markov Model) [70], [53], [86] και παραλλαγές τους που είναι και η πιο διαδεδομένη τεχνική, διάφορα είδη νευρωνικών δικτύων όπως αναδρομικά νευρο-ασαφή δίκτυα (recurrent neural networks) [38] ή νευρωνικά δίκτυα χρονοκαθυστέρησης (time delay neural networks) [92], μηχανές πεπερασμένων καταστάσεων [36] (finite state machines), Bayesian ταξινομητές [88] κ.α. Επίσης έχουν γίνει προσπάθειες συνδυασμού μεθοδολογιών. Στις εργασίες [47], [19] χρησιμοποιήθηκαν

μοντέλα HMM σε συνδυασμό με αυτο-οργανούμενους χάρτες. Στην πρώτη εργασία οι αυτο-οργανούμενου χάρτες χρησιμοποιούνται για την στατική αναγνώριση των χειρονομιών ενώ για την δυναμική αναγνώριση κατά την διάρκεια της εκτέλεσης της χειρονομίας χρησιμοποιήθηκαν μοντέλα HMM. Στην δεύτερη, οι αυτο-οργανούμενοι χάρτες χρησιμοποιήθηκαν ως πρώτο επίπεδο επεξεργασίας ενώ στο δεύτερο επίπεδο έγινε χρήση των μοντέλων HMM.

Η παρούσα εργασία εντάσσεται στον χώρο της μηχανικής μάθησης και ο στόχος είναι η δημιουργία ενός συστήματος που θα εκπαιδεύεται με κατηγοριοποιημένες χειρονομίες, οι οποίες δίνονται με την μορφή ενός συνόλου ζευγών συντεταγμένων που απεικονίζουν την πορεία του χεριού κατά την διάρκεια εκτέλεσης της χειρονομίας και θα έχει την δυνατότητα να αναγνωρίζει στιγμιότυπα χειρονομιών άγνωστης κατηγορίας.

6.1.2 Περιγραφή συστήματος

Κάθε στιγμιότυπο χειρονομίας κωδικοποιείται από μία χρονοσειρά σημείων σε ένα δισδιάστατο χώρο συντεταγμένων. Τα σημεία αυτά αντιστοιχούν στην σχετική θέση του χεριού σε σχέση με την θέση του κεφαλιού κατά την διάρκεια της εκτέλεσης της χειρονομίας. Δηλαδή, ένα στιγμιότυπο G_i μίας χειρονομίας που κωδικοποιείται από l σημεία μπορεί να εκφραστεί ως ένα διατεταγμένο σύνολο σημείων:

$$G_i = \{(x_1, y_1), (x_2, y_2), ...(x_l, y_l)\} \tag{6.1}$$

οπού το l μεταβάλλεται από στιγμιότυπο σε στιγμιότυπο καθώς το πλήθος των σημείων που κωδικοποιούν μία χειρονομία δεν είναι σταθερό. Επομένως, το σύνολο D των δεδομένων εισόδου του συστήματος αποτελείτε από c υποσύνολα D_j όπου το κάθε υποσύνολο περιέχει όλα τα στιγμιότυπα G_i που ανήκουν στην ίδια κατηγορία χειρονομιών.

$$D = \{D_1, D_2, ..., D_c\} : D_j = \{G_1, G_2, ..., G_n\} \tag{6.2}$$

όπου n είναι το πλήθος των στιγμιότυπων των χειρονομιών που ανήκουν στην κατηγορία j.

Η μεθοδολογία βασίζεται στον μετασχηματισμό της κωδικοποίησης της χειρονομίας από μία σειρά σημείων στον χώρο σε μία σειρά συμβόλων που είναι η κατάλληλη μορφή για την δημιουργία πιθανοτικών μοντέλων καταστάσεων. Δύο διαφορετικοί μετασχηματισμοί πραγματοποιούνται από το παρόν σύστημα.

Ο πρώτος βασίζεται στην θέση των σημείων κατά την διάρκεια της χειρονομίας και μετατρέπει τα ζεύγη των συντεταγμένων που ορίζουν αυτά τα σημεία

σε ένα σύνολο από συμβολικές καταστάσεις, οι οποίες αντιπροσωπεύουν διαφορετικές περιοχές του χώρου εισόδου. Αυτές οι καταστάσεις θα χρησιμοποιηθούν για την δημιουργία μοντέλων Markov πρώτης τάξης. Αυτός ο μετασχηματισμός επιτυγχάνεται με την χρήση του μοντέλου των αυτο-οργανούμενων χαρτών όπου κάθε νευρώνας αποτελεί και μία πιθανή κατάσταση. Παρόλο όμως που οι νευρώνες του χάρτη αντιμετωπίζονται σαν συμβολικές καταστάσεις, η θέση των νευρώνων επάνω στο πλέγμα του εκπαιδευμένου χάρτη παρέχει την δυνατότητα ορισμού μιας συνάρτησης απόστασης που θα χρησιμοποιηθεί κατά την διαδικασία κατηγοριοποίησης μίας νέας άγνωστης χειρονομίας.

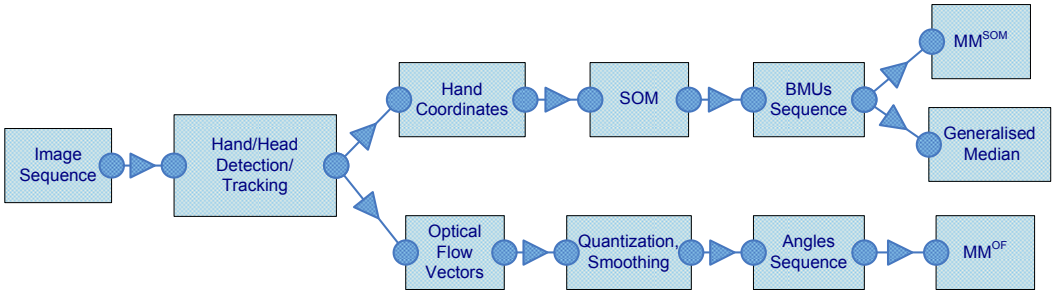
Εφόσον τα στιγμιότυπα των χειρονομιών με την χρήση αυτού του μετασχηματισμού μετατραπούν σε ένα σύνολο από συμβολικές αναπαραστάσεις τότε μπορεί να χρησιμοποιηθεί και μετρική Levenshtein για την σύγκριση μεταξύ δύο διαφορετικών χειρονομιών καθώς και για τον ορισμό μίας «μέσης» χειρονομίας που θα αναπαριστά την μέση τιμή ενός συνόλου χειρονομιών.

Ο δεύτερος μετασχηματισμός βασίζεται στην αναπαράσταση των κατεύθυνσης του χεριού και των αλλαγών της κατά την διάρκεια της χειρονομίας. Ο στόχος είναι να κωδικοποιηθούν οι διαφορετικές κατευθύνσεις έτσι ώστε να είναι δυνατή η απεικόνιση της χειρονομίας με βάση αυτές και να αποτελέσουν μία μορφή αναπαράστασης της χειρονομίας. Ο μετασχηματισμός αυτός δημιουργεί ένα σύνολο κατευθύνσεων-γωνιών που κβαντίζονται σε ένα μικρότερο και εκ των προτέρων, καθορισμένο αριθμό γωνιών, οι οποίες θα αποτελέσουν και τις συμβολικές καταστάσεις που θα χρησιμοποιηθούν για την δημιουργία επιπλέον μοντέλων Markov πρώτης τάξης.

Για την κατηγοριοποίηση άγνωστων χειρονομιών, χρησιμοποιούνται κυρίως τα μοντέλα Markov που δημιουργούνται από τον πρώτο μετασχηματισμό. Τα μοντέλα που δημιουργούνται από τον δεύτερο μετασχηματισμό χρησιμοποιούνται στην περίπτωση που τα προηγούμενα μοντέλα δεν παρουσιάζουν αυξημένο βαθμό αξιοπιστίας στην απόφαση κατηγοριοποίησης. Το σχήμα 6.1 παρουσιάζεται ένα μπλοκ διάγραμμα του συστήματος. Το πρώτο στάδιο αποτελείτε από μία διαδικασία επεξεργασίας εικόνας που παρέχει την δυνατότητα εξαγωγής περιγραφών της μορφής G_i από δεδομένα σε μορφή ακολουθίας εικόνων που απεικονίζουν την εκτέλεση της χειρονομίας [9].

6.1.2.1 Πρώτος μετασχηματισμός

Οι συντεταγμένες των σημείων που χαρακτηρίζουν τις χειρονομίες χρησιμοποιούνται για την εκπαίδευση ενός αυτο-οργανούμενου χάρτη. Κατά την διάρκεια της εκπαίδευσης, οι συντεταγμένες παρουσιάζονται στο μοντέλο σε τυχαία



Σχήμα 6.1: *Μπλοκ διάγραμμα συστήματος.*

σειρά, ανεξάρτητα σε πιο στιγμιότυπο χειρονομίας ανήκουν και με την σειρά με την οποία εμφανίζονται στο εκάστοτε στιγμιότυπο. Το σύστημα χρησιμοποιεί αυτο-οργανούμενο χάρτη με εξαγωνικό πλέγμα και ίσων πλευρικών διαστάσεων. Το μέγεθος του χάρτη δηλαδή ο αριθμός των νευρώνων του πλέγματος καθορίζεται από τον χρήστη. Μετά την εκπαίδευση του χάρτη, κάθε σημείο που έχει χρησιμοποιηθεί στην εκπαίδευση του χάρτη αντιστοιχίζεται με τον νευρώνα νικητή του (BMU), δηλαδή με τον νευρώνα με την μικρότερη απόσταση από το σημείο στον χώρο των δεδομένων εισόδου. Με αυτόν τον τρόπο είναι εφικτός ο μετασχηματισμός της χειρονομίας G_i από μία σειρά από ζεύγη συντεταγμένων σε μία σειρά από νευρώνες.

$$T(G_i) = (u_1, u_2, ..., u_l) : u_i = BMU(x_i, y_i) \tag{6.3}$$

Όπου $T(G_i)$ είναι η μετασχηματισμένη χειρονομία και η συνάρτηση $BMU(x_i, y_i)$ επιστρέφει τον δείκτη νευρώνα νικητή του σημείου με συντεταγμένες (x_i, y_i). Αφού η τιμή u_i είναι ο δείκτης του νικητή νευρώνα, η συνάρτηση αυτή ορίζεται ως $BMU : \Re^2 \rightarrow S$, όπου το S είναι το σύνολο των δεικτών των νευρώνων του εκπαιδευμένου χάρτη και μπορεί να θεωρηθεί και ως ένα σύνολο συμβολικών καταστάσεων.

Στις περισσότερες περιπτώσεις, η τιμή u_i σημείων που είναι διαδοχικά σε μία χειρονομία είναι η ίδια αφού λόγω της συνέχειας στην κίνηση του χεριού κατά την εκτέλεση της χειρονομίας διαδοχικά σημεία εμφανίζονται σχετικά κοντά. Οπότε είναι αναμενόμενο τα σημεία αυτά να ομαδοποιούνται στον ίδιο νευρώνα. Από την αντικατάσταση στην μορφή της χειρονομίας που προκύπτει από την εξίσωση (6.3), των διαδοχικών ίσων τιμών u_i με μία τιμή κάθε φόρα, προκύπτει ο τελικός ορισμός της μετασχηματισμένης χειρονομίας.

$$G_i^{'} = N(T(G_i)) = \{u_1, u_2, ..., u_m, \} : m \leq l, \forall t \in [2, l] u_t \neq u_{t-1} \tag{6.4}$$

όπου $N()$ είναι η συνάρτηση που πραγματοποιεί την αντικατάσταση των ίδιων

διαδοχικών τιμών u_i και $G_i^{'}$ είναι το μετασχηματισμένο στιγμιότυπο της χειρονομίας. Η αφαίρεσή των πολλών επαναλήψεων των ίδιων διαδοχικών τιμών u_i γίνεται γιατί η μετασχηματισμένη χειρονομία $G_i^{'}$ πρέπει να απεικονίζει την πορεία της χειρονομίας στους διαφορετικούς νευρώνες του χάρτη. Ο εκπαιδευμένος χάρτης ομαδοποιεί τα σημεία που συγκροτούν τα διαφορετικά στιγμιότυπα των χειρονομιών καθώς επίσης απεικονίζει την τοπολογία των δεδομένων εισόδου στην θέση των νευρώνων επάνω στο πλέγμα. Καθώς τα δεδομένα εισόδου είναι χωρικές συντεταγμένες, ο χάρτης μπορεί να θεωρηθεί ως ένα μοντέλο χωρικού κβαντισμού του χώρου εισόδου.

Ο μετασχηματισμός αυτός συνολικά απεικονίζει την συνεχή πορεία του χεριού κατά την διάρκεια της χειρονομίας σε ένα σύνολο διακριτών συμβολικών καταστάσεων που είναι οι νευρώνες του εκπαιδευμένου αυτο-οργανούμενου χάρτη και οι οποίοι μπορούν να αποτελέσουν τις καταστάσεις για την δημιουργία μοντέλων Markov πρώτης τάξης.

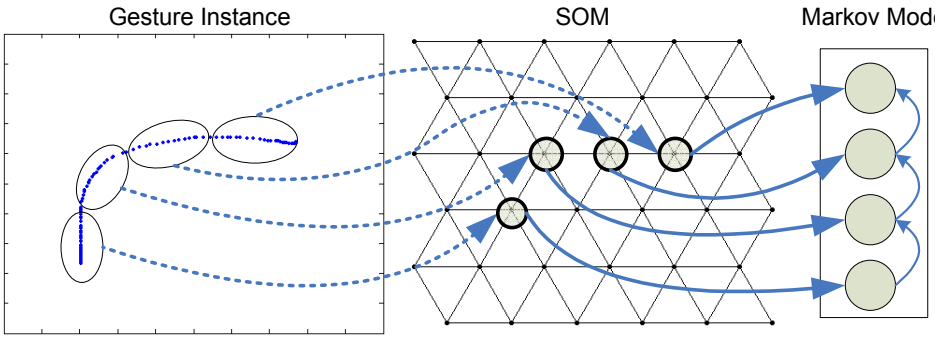
Κάθε σύνολο $D_j^{'}$, δηλαδή το σύνολο που περιέχει όλα τα στιγμιότυπα μετασχηματισμένων χειρονομιών $G_i^{'}$ κατηγορίας j, χρησιμοποιείται για την δημιουργία ενός διαφορετικού μοντέλου Markov. Όπως προαναφέρθηκε, οι καταστάσεις του κάθε μοντέλου είναι οι νευρώνες του χάρτη και η διαδοχή των τιμών u_i σε κάθε μετασχηματισμένη χειρονομία $G_i^{'}$ χρησιμοποιείται για τον υπολογισμό των πιθανοτήτων μεταβάσεις από μία κατάσταση του μοντέλου σε μία άλλη καθώς και για τον υπολογισμό της πιθανότητας της αρχικής κατάστασης του κάθε μοντέλου. Το αποτέλεσμα είναι ένα σύνολο MM^{som} από c μοντέλα Markov, όπου c είναι ο αριθμός των διαφορετικών κατηγοριών χειρονομιών.

$$MM^{som} = \{MM_1^{som}, MM_2^{som}, ..., MM_c^{som}\} : D_i^{'} = \{G_1^{'}, G_2^{'}, ..., G_n^{'}\} \rightarrow MM_i^{som} \tag{6.5}$$

Τα μοντέλα αυτά χρησιμοποιούνται για την ταξινόμηση σε μία από τις c κατηγορίες ενός νέου στιγμιότυπο χειρονομίας άγνωστης κατηγορίας. Το σχήμα 6.2 απεικονίζει συνοπτικά τον μετασχηματισμό με χρήση του αυτο-οργανούμενου χάρτη και την δημιουργία των Markov μοντέλων.

6.1.2.2 Δεύτερος μετασχηματισμός

Με στόχο την καλύτερη περιγραφή της κάθε χειρονομίας, αναπτύχθηκε ένας ακόμα μετασχηματισμός, ο οποίος βασίζεται στην κωδικοποίηση της κατεύθυνσης του χεριού κατά την διάρκεια της χειρονομίας. Δηλαδή κωδικοποιείται η φορά του χεριού και όχι οι συντεταγμένες της θέσης του χεριού. Αυτό επιτυγχάνεται υπολογίζοντας την γωνία που σχηματίζει το διάνυσμα κατεύθυνσης που

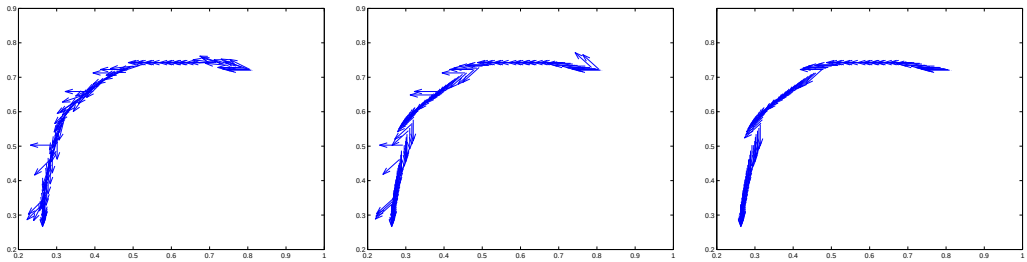


Σχήμα 6.2: *Αντιστοίχιση στιγμιότυπου με τους νευρώνες του αυτο-οργανούμενου χάρτη και δημιουργία μοντέλων Markov*

ορίζεται από δύο διαδοχικά σημεία μίας χειρονομίας σε σχέση με τον άξονα χ του συστήματος συντεταγμένων. Στην συνέχεια αυτές οι γωνίες κβαντίζονται σε οκτώ διαφορετικές τιμές που απεικονίζεται και στο σχήμα 6.4. Για αυτόν τον λόγο ορίζεται η παρακάτω συνάρτηση OF:

$$OF(G_i) = \{v_1, v_2, ..., v_m\} : v_i = W_r\left(Q\left(\arctan\left(\frac{y_i - y_{i-1}}{x_i - x_{i-1}}\right)\right)\right) \qquad (6.6)$$

όπου v_i είναι οι κβαντισμένες τιμές των γωνιών, Q είναι η συνάρτηση κβαντισμού και W_r είναι μία συνάρτηση μέσου που εφαρμόζεται στις τιμές γύρω από την τιμή εισόδου της συνάρτησης, δηλαδή στις γωνίες που προηγούνται και έπονται της τρέχουσας τιμής. Η συνάρτηση αυτή εφαρμόζεται για να εξομαλύνει τις τιμές των γωνιών μετά τον κβαντισμό τους. Η εξομάλυνση αυτή είναι απαραίτητη διότι κατά την εκτέλεση της χειρονομίας δημιουργούνται πολύ σύντομες αλλαγές κατεύθυνσης, οι οποίες όμως δεν οφείλονται σε πραγματικές αλλαγές κατεύθυνσης αλλά στη αστάθεια του χεριού κατά την εκτέλεση της χειρονομίας (σχήμα).



(α) (β) (γ)

Σχήμα 6.3: *Διανύσματα κατεύθυνσης στιγμιότυπου χειρονομίας. (α)Αρχικά διανύσματα κατεύθυνσης. (β) Διανύσματα κβαντισμένων κατευθύνσεων. (γ) Διανύσματα εξομαλισμένων κατευθύνσεων.*

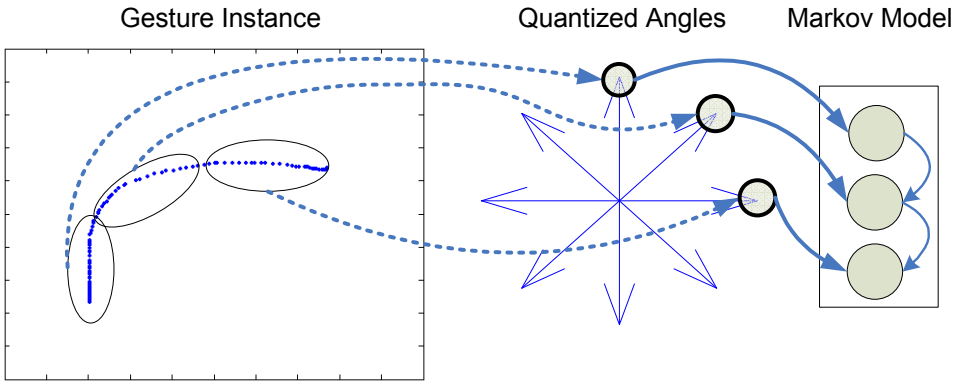
Από την εφαρμογή του παραπάνω μετασχηματισμού σε συνδυασμό με την εφαρμογή της συνάρτησης $N()$ (εξ. 6.4) για την αφαίρεση διαδοχικών ίδιων τιμών προκύπτει η νέα μετασχηματισμένη χειρονομία:

$$G_i^{''} = N(OF(G_i)) = \{v_1, v_2, ..., v_m\} \tag{6.7}$$

Οι τιμές v_i ορίζουν και τις καταστάσεις ενός νέου συνόλου από μοντέλα Markov MM^{of} τα οποία δημιουργούνται χρησιμοποιώντας τα νέα σύνολα μετασχηματισμένων χειρονομιών $D_i^{''}$.

$$MM^{of} = \{MM_1^{of}, MM_2^{of}, ..., MM_c^{of}\} : D_i^{''} = \{G_1^{''}, G_2^{''}, ..., G_n^{''}\} \rightarrow MM_i^{of} \tag{6.8}$$

Το σχήμα 6.4 απεικονίζει τον μετασχηματισμό της χειρονομίας σε ένα σύνολο κβαντισμένων γωνιών από όπου προκύπτουν τα νέα μοντέλα Markov.



Σχήμα 6.4: *Αντιστοίχιση στιγμιότυπου χειρονομίας σε κβαντισμένες κατευθύνσεις και δημιουργία μοντέλων Markov*

6.1.2.3 Κατηγοριοποίηση άγνωστης χειρονομίας

Μετά την εκπαίδευση του αυτο-οργανούμενου χάρτη και την δημιουργία των μοντέλων Markov, το σύστημα είναι σε θέση να κατηγοριοποιήσει στιγμιότυπα χειρονομιών άγνωστης κατηγορίας. Έστω G_k το στιγμιότυπο μίας χειρονομίας άγνωστης κατηγορίας. Με χρήση των δύο μετασχηματισμών (εξ. 6.4 και 6.7) προκύπτουν τα μετασχηματισμένα στιγμιότυπα $G_k^{'}$ και $G_k^{''}$.

$$G_k^{'} = \{u_1, u_2, ..., u_m\} \tag{6.9}$$

$$G_k^{''} = \{v_1, v_2, ..., v_q\} \tag{6.10}$$

Χρησιμοποιώντας τα μοντέλα MM^{som}, υπολογίζεται ένας βαθμός αξιολόγησης του άγνωστου στιγμιότυπου ως προς την κατηγορία j ως εξής:

$$P(G_k^{'}|MM_j^{som}) = \frac{\sum_{i=1}^{m} S_i^{som}}{m} \tag{6.11}$$

Όπου m είναι το πλήθος των στοιχείων του $G_k^{'}$. Η παραπάνω εξίσωση υπολογίσει την μέση τιμή του παράγοντα S_i^{som}, ο οποίος αντιπροσωπεύει ένα βαθμό αξιολόγησης της τιμής u_i σε σχέση με το μοντέλο MM_j^{som}. Ο παράγοντας αυτός υπολογίζεται ως:

$$S_i^{som} = \max_z \left({NF_{u_i}}^{som}(z)\, P\left(z|u_{i-1}, MM_j^{som}\right)\right) \tag{6.12}$$

όπου z είναι μία μεταβλητή που παίρνει τιμές ίσες με τους δείκτες τους νευρώνων του εκπαιδευμένου χάρτη, δηλαδή όλες τις πιθανές καταστάσεις του μοντέλου MM_j^{som}. Η συνάρτηση $NF_{u_i}^{som}(z)$ επιστρέφει την απόσταση του z με βάση την Gaussian συνάρτησης γειτνίασης του αυτο-οργανούμενου χάρτη με κέντρο τον νευρώνα u_i. Η τιμή u_i κανονικά δίνεται από το μετασχηματισμένο στιγμιότυπο $G_k^{'}$, όμως κατά τον υπολογισμό της τιμής S_i^{som} χρησιμοποιείται η παρακάτω τιμή:

$$u_i = \arg\max_z (S_i^{som}) \tag{6.13}$$

Στην εξίσωση 6.12, η γειτνίαση του νευρώνα z και της τρέχουσας κατάστασης-νευρώνα u_i πολλαπλασιάζεται με την πιθανότητα μετάβασης από την προηγούμενη κατάσταση-νευρώνα u_{i-1} στην κατάσταση-νευρώνα z που προκύπτει από τον πίνακα μεταβάσεων του μοντέλου MM_j^{som}. Καθώς η τιμή z μεταβάλλεται, από την μέγιστη τιμή του γινομένου αυτού θα προκύψει μία κατάσταση που θα συνδυάζει αυξημένη πιθανότητα μετάβασης από την προηγούμενη κατάσταση με μειωμένη απόσταση από την τρέχουσα κατάσταση. Δηλαδή γίνεται μία αναζήτηση ανάμεσα στους γειτονικούς νευρώνες της τρέχουσας κατάστασης για τον νευρώνα με την σχετικά μεγαλύτερη πιθανότητα μετάβασης από τον προηγούμενο. Ο νευρώνας που θα προκύψει χρησιμοποιείται και ως προηγούμενη κατάσταση καθώς η αξιολόγηση προχωρεί στην επόμενη τιμή u_i, όπως προκύπτει και από την εξίσωση 6.13. Η αρχική τιμή S_1 για την εξίσωση 6.12 προκύπτει ως:

$$S_1^{som} = \max_z \left(NF_{u_1}^{som}(z) \pi_j^{som}(z)\right) \tag{6.14}$$

όπου $\pi_j^{som}(z)$ είναι η πιθανότητα να είναι η κατάσταση z η αρχική κατάσταση του μοντέλου MM_j^{som}.

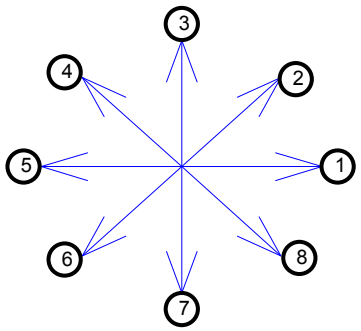
Με τον ίδιο τρόπο αλλά χρησιμοποιώντας τα μοντέλα MM^{of}, υπολογίζεται ένας βαθμός αξιολόγησης του άγνωστου στιγμιότυπου ως προς την κατηγορία j με τις αντίστοιχες εξισώσεις:

$$P(G_k''|MM_j^{of}) = \frac{\sum_{i=1}^{q} S_i^{of}}{q} \tag{6.15}$$

Όπου q είναι το πλήθος των στοιχείων του G_k''. Η τιμή S_i^{of} υπολογίζεται ως:

$$S_i^{of} = \max_z \left({NF_{v_i}}^{of}(z)\, P\left(z|v_{i-1}, MM_j^{of}\right)\right) \tag{6.16}$$

Η διαφορά είναι ότι σε αυτήν την περίπτωση, οι πιθανές καταστάσεις είναι οι διαφορετικές τιμές των κβαντισμένων γωνιών και η συνάρτηση γειτνίασης $NF_{v_i}^{of}(z)$ υπολογίζεται από ένα πολύ απλό πίνακα γειτνίασης μεταξύ των διαφορετικών πιθανών γωνιών όπως φαίνεται στο σχήμα 6.5.



angles labels v_i

angles labels v_i	1	2	3	4	5	6	7	8
1	1	0.5	0.25	0	0	0	0.25	0.5
2	0.5	1	0.5	0.25	0	0	0	0.25
3	0.25	0.5	1	0.5	0.25	0	0	0
4	0	0.25	0.5	1	0.5	0.25	0	0
5	0	0	0.25	0.5	1	0.5	0.25	0
6	0	0	0	0.25	0.5	1	0.5	0.25
7	0.25	0	0	0	0.25	0.5	1	0.5
8	0.5	0.25	0	0	0	0.25	0.5	1

Σχήμα 6.5: *Συμβολισμός και πίνακας γειτνίασης κβαντισμένων διανυσμάτων κατεύθυνσης.*

Η τιμή v_i υπολογίζεται ομοίως

$$v_i = \arg\max_z \left(S_i^{of} \right) \tag{6.17}$$

και η αρχική τιμή S_1^{of} ως:

$$S_1^{of} = \max_z \left(NF_{v_1}^{of}(z) \pi_j^{of}(z) \right) \tag{6.18}$$

Μέσος όρος κατηγορίας

Στο παρόν σύστημα κρίθηκε απαραίτητο να πραγματοποιείται σύγκριση και του μήκους της κάθε άγνωστης χειρονομίας με το μήκος των χειρονομιών της

κάθε κατηγορίας και καθώς αυτή η σύγκριση δεν μπορεί να πραγματοποιηθεί με χρήση των μοντέλων Markov ορίστηκε ένα πρωτότυπο χειρονομίας για κάθε χειρονομία. Αυτό το πρωτότυπο είναι μία χειρονομία αντιπροσωπεύει την «μέση τιμή» του κάθε συνόλου $D_j^{'}$. Επειδή όμως το σύνολο αυτό είναι ένα σύνολο συμβολοσειρών, η θεωρούμενη ως μέση τιμή ορίζεται ως εξής. Έστω το S είναι ένα σύνολο από συμβολοσειρές s_i. Ως μέση τιμή μπορεί να οριστεί μία συμβολοσειρά m, η οποία θα ελαχιστοποιεί την ακόλουθη παράσταση.

$$\sum_{s_i} L\left(s_i, m\right), \forall s_i \in S \tag{6.19}$$

Η συνάρτηση $L()$ αντιστοιχεί στην απόσταση Levenshtein [45], η οποία είναι η πιο διαδεδομένη μετρική για την σύγκριση μεταξύ συμβολοσειρών. Εάν τεθεί ο περιορισμός η συμβολοσειρά m να είναι ένα από τα στοιχεία του συνόλου S, τότε αυτή μπορεί να θεωρηθεί ως ο «μέσος» του συνόλου S. Εάν όμως είναι μία υποθετική συμβολοσειρά και η αναζήτηση δεν περιοριστεί στα στοιχεία του S τότε το m αντιστοιχεί στην «μέση τιμή» του συνόλου S. Η διαδικασία προσδιορισμού του μπορεί να γίνει αφού πρώτα βρεθεί ο «μέσος», ο οποίος προκύπτει με την επιλογή του στοιχείου του συνόλου S με το μικρότερο άθροισμα αποστάσεων από όλα τα υπόλοιπα στοιχεία του συνόλου. Στην συνέχεια η «μέση τιμή» υπολογίζεται εάν σε κάθε ένα από τα σύμβολα του «μέσου» εφαρμοστούν οι βασικές ενέργειες της απόστασης Levenshtein, δηλαδή εισαγωγή, αντιγραφή, αντικατάσταση και η κάθε ενέργεια γίνεται αποδεκτή εφόσον το νέο άθροισμα των αποστάσεων από όλα τα στοιχεία του συνόλου ελαττώνεται [41].

Με χρήση του παραπάνω ορισμού υπολογίζεται η «μέση τιμή» $M(D_j^{'})$ του συνόλου $D_j^{'}$ και ορίζεται και η απόσταση $L_{kj} = L\left(G_k^{'}|M(D_j^{'})\right)$ μεταξύ αυτής της μέσης τιμής και της άγνωστης χειρονομίας $G_k^{'}$.

Κανόνας κατηγοριοποίησης

Η κατηγορία ενός νέου στιγμιότυπου εκτιμάται κυρίως με βάση τα μοντέλα MM^{som}, οπότε η κατηγορία του νέου στιγμιότυπου προκύπτει από την σχέση:

$$\arg\max_j \left(P\left(G_k^{'} \left| MM_j^{som}\right.\right)\right) \tag{6.20}$$

Όμως για να ληφθεί η απόφαση κατηγοριοποίησης με βάση την παραπάνω εξίσωση πρέπει να ισχύουν όλες οι παρακάτω συνθήκες.

$$\max_j \left(P\left(G_k^{'}|MM_j^{som}\right)\right) \geq \alpha \tag{6.21}$$

$$\max_j \left(P\left(G_k^{'} | MM_j^{som} \right) \right) - 2^{nd} \max_j \left(P\left(G_k^{'} | MM_j^{som} \right) \right) \geq \beta \tag{6.22}$$

$$L_{k,\arg\max_j \left(P\left(G_k^{'} | MM_j^{som} \right) \right)} \leq \gamma LM \left(\arg\max_j \left(P\left(G_k^{'} | MM_j^{som} \right) \right) \right) \tag{6.23}$$

Οι δύο πρώτες συνθήκες απαιτούν ότι η μέγιστη τιμή του βαθμού αξιολόγησης, πρώτον να είναι μεγαλύτερη από μία σταθερή παράμετρο α και δεύτερον η διαφορά από την αμέσως μικρότερη να επίσης μεγαλύτερη μια σταθερής παραμέτρου β. Οι δύο αυτές παράμετροι αντιπροσωπεύουν στάθμες εμπιστοσύνης, οι οποίες πρέπει να υπερβαίνονται έτσι ώστε να θεωρείτε η απόφαση κατηγοριοποίησης αξιόπιστη. Η τρίτη συνθήκη εισάγεται έτσι ώστε να γίνει σύγκριση με βάση την απόσταση Levenshtein μεταξύ της άγνωστης χειρονομίας και την «μέση τιμή» του συνόλου $D_j^{'}$ με τον μέγιστο βαθμό αξιολόγησης. Η απόσταση αυτή δεν πρέπει να μεγαλύτερη από την τιμή της συνάρτησης $LM(j)$ για την κατηγορία με τον μεγαλύτερο βαθμό αξιολόγησης, πολλαπλασιαζόμενη με μία παράμετρο γ. Η συνάρτηση αυτή υπολογίζει την μέση τιμή των αποστάσεων Levenshtein μεταξύ της «μέσης» χειρονομίας και όλων των στιγμιοτύπων των χειρονομιών του συνόλου $D_j^{'}$. Η τελευταία συνθήκη έχει ως στόχο να αξιολογήσει την χειρονομία με βάση την διαφορά της από μία μέση έκφραση του συνόλου $D_j^{'}$.

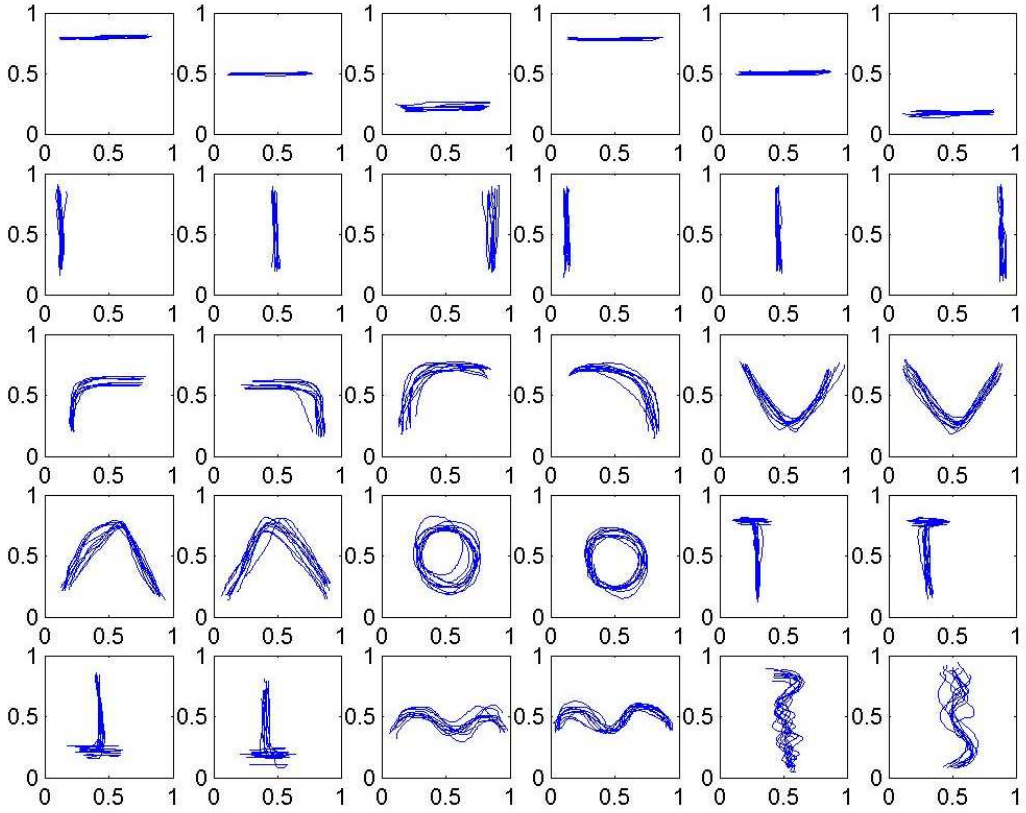
Εάν δεν πληρούνται όλες οι παραπάνω συνθήκες τότε η τελική απόφαση της κατηγοριοποίησης λαμβάνεται με βάση την παρακάτω εξίσωση.

$$\arg\max_j \left(P\left(G_k^{'} | MM_j^{som} \right) P\left(G_k^{''} | MM_j^{of} \right) \frac{1}{\frac{L_{kj}}{\|M(D_j)\|}} \right) \tag{6.24}$$

Η έκφραση αυτή συνδυάζει τον βαθμό αξιολόγησης που προκύπτει και από τα δύο σύνολα μοντέλων Markov MM^{som} και MM^{of} καθώς και τον λόγο της απόστασης Levenshtein μεταξύ της χειρονομίας και της «μέσης» χειρονομίας της κάθε κατηγορίας ως προς το μήκος αυτής της «μέσης» χειρονομίας.

6.1.3 Πειραματική αξιολόγηση

Η πειραματική αξιολόγηση του συστήματος έγινε με την χρήση ενός συνόλου δεδομένων 30 διαφορετικών κατηγοριών χειρονομιών. Κάθε κατηγορία χειρονομίας περιγράφεται από 10 στιγμιότυπα-επαναλήψεις της χειρονομίας. Με την



Σχήμα 6.6: *Σύνολο δεδομένων των 30 διαφορετικών κατηγοριών.*

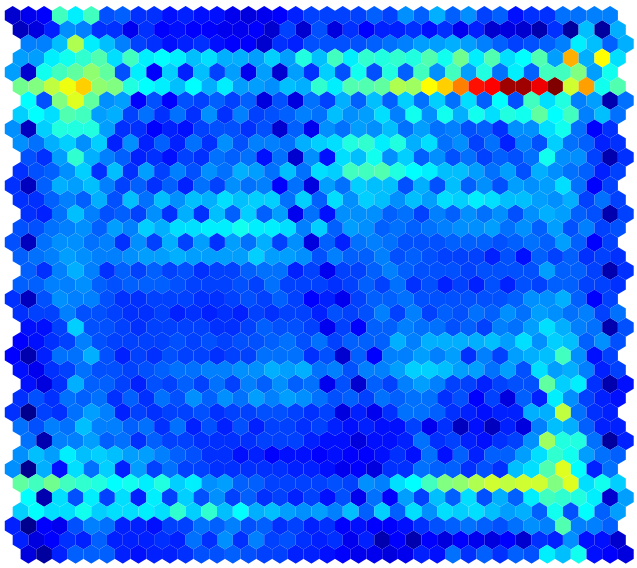
χρήση του υποσυστήματος επεξεργασίας εικόνων γίνεται η εξαγωγή των περιγραφών G_i των χειρονομιών. Στο σχήμα 6.6 απεικονίζονται όλες οι κατηγορίες και όλα τα στιγμιότυπα της κάθε κατηγορίας.

Η πειραματική ανάλυση έγινε με την χρήση αυτού του συνόλου δεδομένων με σκοπό την αξιολόγηση της ικανότητας κατηγοριοποίησης του συστήματος. Με χρήση όλου του συνόλου δεδομένων για την εκπαίδευση του συστήματος και για τον έλεγχο της ορθής κατηγοριοποίησης προκύπτει ότι το σύστημα επιτυγχάνει ποσοστό 100% ορθής κατηγοριοποίησης. Το αποτέλεσμα αυτό είναι ενδεικτικό της ισχυρής ικανότητας μάθησης του συστήματος. Με σκοπό να αξιολογηθεί και ικανότητα γενίκευσης του συστήματος σε στιγμιότυπα χειρονομιών τα οποία δεν περιλαμβάνονται στο σύνολο εκπαίδευσης, πραγματοποιήθηκε πειραματική αξιολόγηση με χρήση της στρατηγικής 10-fold cross validation. Σε αυτήν την περίπτωση το συνολικό ποσοστό ορθής κατηγοριοποίησης προέκυψε ίσο με 93%. Τα αναλυτικά αποτελέσματα με τα ποσοστά ορθής κατηγοριοποίησης για κάθε κατηγορία περιέχονται στον πίνακα 6.1.

Στο σχήμα 6.7 απεικονίζεται και η διάταξη U-matrix του αυτο-οργανούμενου χάρτη που προέκυψε από την εκπαίδευση του συστήματος με ολόκληρο το σύ-

Πίνακας 6.1: *Αποτελέσματα ορθής κατηγοριοποίησης ανά κατηγορία.*

Κατηγορία	(%)	Κατηγορία	(%)	Κατηγορία	(%)	Κατηγορία	(%)	Κατηγορία	(%)
1	100	7	100	13	80	19	100	25	90
2	100	8	100	14	80	20	90	26	90
3	100	9	100	15	100	21	50	27	90
4	100	10	100	16	90	22	70	28	100
5	100	11	100	17	100	23	100	29	100
6	100	12	100	18	100	24	60	30	100



Σχήμα 6.7: *Σύνολο δεδομένων των 30 διαφορετικών κατηγοριών.*

νολο δεδομένων. Ο χρωματισμός είναι ανάλογος της Ευκλείδειας απόστασης μεταξύ των συνδεδεμένων νευρώνων του πλέγματος. Η απόσταση υπολογίζεται χρησιμοποιώντας τα διανύσματα των νευρώνων στον χώρο εισόδου. Οι αποχρώσεις του μπλε χρώματος αντιστοιχούν στις μικρότερες αποστάσεις ενώ αντίθετα οι αποχρώσεις του κόκκινου αντιστοιχούν στις μεγαλύτερες αποστάσεις. Οι ενδιάμεσοι χρωματισμοί αντιστοιχούν σε ανάλογες αποστάσεις. Όπως φαίνεται και από το σχήμα 6.6, τα σημεία που αποτελούν τα στιγμιότυπα των χειρονομιών καλύπτουν σχεδόν όλο τον χώρο εισόδου λόγω της ποικιλίας των χειρονομιών και στην μη απόλυτη ομοιότητα μεταξύ των στιγμιότυπων της ίδιας κατηγορίας. Με βάση αυτήν την παρατήρηση, ο εκπαιδευμένος χάρτης παρουσιάζει, όπως αναμενόταν άλλωστε, ομοιόμορφη κατανομή των νευρώνων στο μεγαλύτερο μέρος του, γεγονός που είναι επιθυμητό λόγω του ρόλου του χάρτη στην λειτουργία του συστήματος.

Πίνακας 6.2: *Αποτελέσματα συστήματος HMM ορθής κατηγοριοποίησης ανά κατηγορία.*

Κατηγορία (%)		Κατηγορία (%)		Κατηγορία (%)		Κατηγορία (%)		Κατηγορία (%)	
1	100	7	100	13	100	19	60	25	100
2	100	8	60	14	90	20	100	26	40
3	100	9	70	15	90	21	10	27	90
4	100	10	70	16	100	22	80	28	100
5	100	11	100	17	100	23	100	29	100
6	100	12	100	18	10	24	80	30	100

Επίσης πραγματοποιήθηκαν πειράματα με το ίδιο σύνολο δεδομένων χρησιμοποιώντας ένα σύστημα κατηγοριοποίησης βασισμένο σε μοντέλα HMM (Hidden Markov Model) [39]. Όπως και πριν, εφαρμόστηκε στρατηγική 10-fold cross validation. Χρησιμοποιήθηκε ένα μοντέλο HMM τύπου Bakis για κάθε κατηγορία, το οποίο οριζόταν με χρήση μείγματος τριών Gaussian συναρτήσεων πυκνότητας πιθανότητας. Το αποτέλεσμα αυτού του πειράματος ήταν ένα ποσοστό ορθών κατηγοριοποιήσεων ίσο με 85%. Στον πίνακα 6.2 παρουσιάζονται αναλυτικά τα αποτελέσματα ανά κατηγορία χειρονομίας.

6.2 Εξαγωγή κανόνων από αυτο-οργανούμενους χάρτες

6.2.1 Εισαγωγή

Τα νευρωνικά δίκτυα λειτουργούν αξιόπιστα και με ικανοποιητικά αποτελέσματα σε προβλήματα ομαδοποίηση, ταξινόμησης, προσέγγισης συναρτήσεων κ.α.. Η ικανότητα των νευρωνικών δικτύων να αντιμετωπίζουν αυτού του είδους τα προβλήματα βρίσκεται στην γνώση που αποκομίζουν από τα δεδομένα κατά την διάρκεια της εκπαίδευσής τους από τα δεδομένα εισόδου. Η γνώση που εξάγεται από τα δεδομένα αυτά και χρησιμοποιείται για την επίλυση των προβλημάτων βρίσκεται στην δομή του δικτύου, δηλαδή στο σύνολο των εσωτερικών παραμέτρων του δικτύου που ρυθμίζονται κατά την διάρκεια της εκπαίδευσης. Ο τρόπος αυτός όμως αποθήκευσης της εξαγόμενης γνώσης αλλά και ο τρόπος χρήσης της είναι δύσκολο να διατυπωθεί σε μορφή που να είναι εύκολα κατανοητή και χρηστική για τον άνθρωπο. Σε πολλές εφαρμογές εξόρυξης γνώσης η χρήση νευρωνικών δικτύων και η πρόβλεψη που αυτά προσφέρουν δεν είναι

αρκετή αφού η κατανόηση των δεδομένων είναι πολύ σημαντική.

Μία μορφή που είναι ιδιαίτερα απλή και χρηστική για τη συμβολική παράσταση της εξαγόμενης γνώσης είναι η προτασιακή λογική (propositional logic). Τα νευρωνικά δίκτυα έχουν διαφορετικές δυνατότητες αφού εκτελούν υποσυμβολική επεξεργασία δεδομένων και πολλές φορές είναι πιο ισχυρά από ένα σύνολο λογικών κανόνων. Επίσης γνώσεις που υπάρχουν για κάποιο πρόβλημα βρίσκονται σε συμβολική μορφή και συνήθως είναι δύσκολο να συνδυαστούν με τα νευρωνικά δίκτυα. Η εξαγωγή γνώσης από νευρωνικά δίκτυα, ο συνδυασμός της με τη συμβολική γνώση που ήδη υπάρχει και η χρήση των κανόνων σε έμπειρα συστήματα μπορεί να δημιουργήσει συστήματα αυξημένης αξιοπιστίας που θα μπορούν να ανακαλύπτουν σημαντικές αλληλεπιδράσεις των δεδομένων και πληροφορίες για το υπό εξέταση πρόβλημα.

Ειδικότερα, η χρήση αυτο-οργανούμενων χαρτών για την εξόρυξη γνώσης προσφέρει ομαδοποίηση των δεδομένων και αποδοτική οπτικοποίηση των αποτελεσμάτων ανεξάρτητα του αριθμού των διαστάσεων των δεδομένων. Αν τα πλεονεκτήματα αυτά συνδυαστούν με ένα σύστημα εξαγωγής κανόνων η περιγραφή των δεδομένων μέσω ενός αυτο-οργανούμενου χάρτη θα είναι ακόμα πιο πλήρης. Τα συστήματα που έχουν σχεδιαστεί για να εξάγουν γνώση σε μορφή κανόνων από νευρωνικά δίκτυα βασίζονται κυρίως σε δίκτυα πρόσθιας τροφοδότησης (feed-forward neural networks) [14] παρόλο που η χρήση αυτο-οργανούμενων χαρτών είναι μια επίσης δημοφιλής τεχνική σε εφαρμογές εξόρυξης γνώσης από δεδομένα.

6.2.2 Κανόνες

Η δομή των κανόνων της προτασιακής λογικής είναι τέτοια που πάντα δημιουργούνται σύνορα αποφάσεων στον πολυδιάστατο χώρο του συνόλου δεδομένων. Η γενική μορφή των κανόνων είναι:

$$\text{if } \mathbf{x} \in X^i \text{ then } \text{Κατηγορία}(\mathbf{x}) = C_k \tag{6.25}$$

δηλαδή αν το πρότυπο $\mathbf{x}$ ανήκει στο υποσύνολο X^i του χώρου του συνόλου δεδομένων τότε πρέπει να κατηγοριοποιηθεί στην κατηγορία C_k. Η συνθήκη του κανόνα μπορεί να είναι μια λογική πρόταση του τύπου $L(x_i)$, όπου x_i είναι ένα διάστημα τιμών της μεταβλητής, ή ένας συνδυασμός λογικών προτάσεων. Ο συνδυασμός μπορεί να γίνει είτε με σύζευξη (λογικό «ΚΑΙ»), είτε με διάζευξη (λογικό «Η»), είτε με συνδυασμό συζεύξεων και διαζεύξεων οπότε ο κανόνας γίνεται:

$$\text{If } (L_1(x_1) \wedge L_2(x_2) \wedge ... \wedge L_m(x_m)) \vee L_{m+1}(x_{m+1}) \vee L_{m+2}(x_{m+2}) \vee ... \vee L_n(x_n) \\ \text{then Category}(\mathbf{x}) = \mathbf{C_k} \tag{6.26}$$

Τα σύνορα αποφάσεων που δημιουργεί ένα σύνολο κανόνων δεν μπορούν να προσεγγίσουν καλά τα αποτελέσματα των νευρωνικών δικτύων αν υπάρχουν μόνο λίγοι κανόνες. Η ακρίβεια του συνόλου κανόνων βελτιώνεται με την εισαγωγή περισσότερων κανόνων, κάτι που όμως οδηγεί σε λιγότερο κατανοητή περιγραφή. Αναγκαστικά σε ένα σύστημα κανόνων θα πρέπει να υπάρχει ένας συμβιβασμός μεταξύ απλότητας και ακρίβειας.

6.2.2.1 Μέτρα σημαντικότητας κανόνων

Σκοπός των μέτρων σημαντικότητας κανόνων είναι να προσφέρουν μια αξιολόγηση των κανόνων και να βάζουν ένα όριο στο μεγάλο αριθμό κανόνων που δημιουργούνται από τους αλγορίθμους. Υπάρχουν δύο βασικοί τρόποι για την αξιολόγηση των κανόνων, ο αντικειμενικός και ο υποκειμενικός [50].

Η αντικειμενική προσέγγιση βασίζεται στη δομή και την ακρίβεια των κανόνων αλλά μερικές φορές δεν μπορεί να περιγράψει πλήρως την πολυπλοκότητα των δεδομένων. Η υποκειμενική προσέγγιση βασίζεται και στον χρήστη ή τον ειδικό που εξετάζει τα αποτελέσματα. Η ανάγκη για την ύπαρξη υποκειμενικής προσέγγισης οφείλεται στο γεγονός ότι μόνο έτσι μπορεί να βρεθεί απροσδόκητη πληροφορία σχετικά με τα δεδομένα που δεν ήταν γνωστή στους ειδικούς.

Η αντικειμενική προσέγγιση προσφέρει ποσοτικές πληροφορίες για την εγκυρότητα και την σημασία των κανόνων. Ένας κανόνα R_i μπορεί να θεωρηθεί με δυο διαφορετικούς τρόπους, ως χαρακτηριστικός (characterizing) και ως διαχωριστικός (differentiating) κανόνας [81]. Ο χαρακτηριστικός κανόνας δηλώνει ότι αν ένα πρότυπο ανήκει στην κατηγορία C_i τότε θα ικανοποιεί και τον αντίστοιχο κανόνα δηλαδή:

$$R_i^c : \quad \mathbf{x} \in C_i \Rightarrow x_k \in [\alpha_k, \beta_k] \tag{6.27}$$

Ο διαχωριστικός κανόνας δηλώνει ότι αν ένα πρότυπο ικανοποιεί τη συνθήκη του κανόνα που έχει κατασκευαστεί για μια κατηγορία τότε θα ανήκει στην κατηγορία αυτή δηλαδή:

$$R_i^d : \quad x_k \in [\alpha_k, \beta_k] \Rightarrow \mathbf{x} \in C_i \tag{6.28}$$

Το σύνολο $[\alpha_k, \beta_k]$ υποδηλώνει τα όρια μέσα στα οποία πρέπει να βρίσκεται το χαρακτηριστικό k του προτύπου. Η εγκυρότητα των παραπάνω κανόνων

ορίζεται με ανάλογο τρόπο ως:

$$\begin{aligned} P_i^c &= P(x_k \in [\alpha_k, \beta_k] | C_i) \\ P_i^d &= P(C_i | x_k \in [\alpha_k, \beta_k]) \end{aligned} \tag{6.29}$$

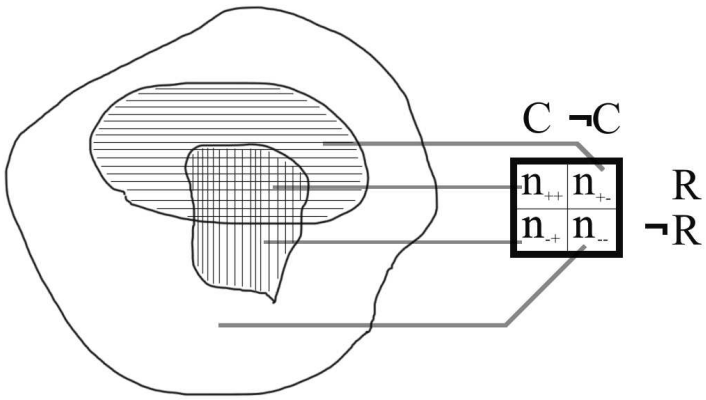
Για την περαιτέρω ανάλυση των μέτρων αξιοπιστίας έστω ένα σύνολο δεδομένων n προτύπων όπου η κατηγορία C_i έχει n_c πρότυπα. Από την εφαρμογή ενός κανόνα σε ένα σύνολο δεδομένων προκύπτουν οι παρακάτω ποσότητες οι οποίες εξηγούνται και γραφικά στο σχήμα 6.8.

1. n_{++} είναι τα πρότυπα που επαληθεύουν τον κανόνα και ανήκουν στη κατηγορία C_i.

2. n_{+-} είναι τα πρότυπα που επαληθεύουν τον κανόνα και δεν ανήκουν στη κατηγορία C_i.

3. $n_{-+} = n_c - n_{++}$ είναι τα πρότυπα που δεν επαληθεύουν τον κανόνα αλλά ανήκουν στη κατηγορία C_i.

4. $n_{--} = n - n_{++} - n_{+-} - n_{-+} = n - n_c - n_{+-}$ είναι τα πρότυπα που ούτε επαληθεύουν τον κανόνα και ούτε ανήκουν στη κατηγορία C_i.



Σχήμα 6.8: *Σχέση ανάμεσα στον κανόνα R και την κατηγορία C. Η κατηγορία C βρίσκεται στην κάθετα σκιασμένη περιοχή και ο κανόνας R στην οριζόντια σκιασμένη περιοχή.*

Με βάση τις παραπάνω ποσότητες μπορούν να οριστούν πολλά διαφορετικά μέτρα αξιολόγησης τα οποία ανάλογα με την έκφραση τους μεγιστοποιούνται ή ελαχιστοποιούνται όταν ο κανόνας περιγράφει τέλεια την κατηγορία και δεν κατηγοριοποιεί λάθος κανένα πρότυπο. Στην ιδανική περίπτωση είναι $n_{+-} = n_{-+} = 0$. Παρακάτω αναφέρονται μερικά από τα κριτήρια που μπορούν να οριστούν.

$$
\begin{aligned}
S_1 &= \frac{n_{++}+n_{--}}{n_{++}+n_{+-}+n_{-+}+n_{--}} \\
S_2 &= P_i^d \cdot P_i^C = \frac{n_{++}}{n_{++}+n_{+-}} \cdot \frac{n_{++}}{n_{++}+n_{-+}} \\
S_3 &= \frac{n_{++}}{n_{++}+n_{+-}+n_{-+}} \\
S_4 &= \frac{n_{++}}{n_C} \\
S_5 &= \frac{n_{++}}{n_{+-}} \\
S_6 &= n_{+-} + n_{-+} = n_{+-} + n_C - n_{++}
\end{aligned}
\tag{6.30}
$$

Το κριτήριο S_1 παρουσιάζει το μειονέκτημα ότι αν τα πρότυπα μιας κατηγορίας είναι πολύ λιγότερα από όλα τα πρότυπα ($n_c \ll n$) τότε η τιμή του κριτηρίου επηρεάζεται πολύ από την ανάγκη να κατηγοριοποιηθούν τα περισσότερα πρότυπα ως n_{--} και δεν μπορεί να πάρει μεγάλο εύρος τιμών. Επίσης τα n_{--} πρότυπα δεν έχουν κάποιο ουσιαστικό ενδιαφέρον όταν εξετάζεται η σχέση ανάμεσα στον κανόνα R_i και την κατηγορία C_i. Για αυτούς τους λόγους το S_1 δεν χρησιμοποιείται στην πράξη.

Το κριτήριο S_2 είναι το γινόμενο των πιθανοτήτων P_i^c και P_i^d και μπορεί να θεωρηθεί ως μέτρο αμοιβαίας εμπιστοσύνης. Τα κριτήρια S_3, S_4 και S_5 είναι απλούστερες παραλλαγές του S_2 και το S_3 προσεγγίζει το S_2 για $n_{++} \gg n_{+-} + n_{-+}$. Το κριτήριο $S6$ είναι μέτρο του πλήθους των προτύπων που κατηγοριοποιούνται λάθος.

Υπάρχουν περιπτώσεις που οι συνθήκες του συνόλου κανόνων δημιουργούν επικαλυπτόμενες περιοχές στο χώρο των δεδομένων και υπάρχουν πρότυπα που επαληθεύουν παραπάνω από έναν κανόνα. Για να λυθεί το πρόβλημα πρέπει να επιλεγεί για αυτά τα πρότυπα σε ποια από τις κατηγορίες θα πρέπει να ενταχθούν. Η επιλογή μπορεί να γίνει με διαφορετικούς τρόπους όπως με τυχαίο τρόπο, να επιλεγεί ο πρώτος κανόνας που επαληθεύεται, να βαθμολογηθούν οι κανόνες ανάλογα με τα πρότυπα που κατηγοριοποιούν ή να επιλεγεί ο κανόνας με τη μικρότερη πιθανότητα λάθους [20]. Η πιθανότητα λάθους (false positive rate) ορίζεται ως:

$$FP = \frac{n_{+-}}{n - n_c} \tag{6.31}$$

και χρησιμοποιείται ώστε να υπάρχει η μικρότερη δυνατή πιθανότητα για λάθος κατηγοριοποίηση προτύπων.

Η ακρίβεια ενός συνόλου κανόνων μπορεί να αυξηθεί κάνοντας πιο αυστηρές τις συνθήκες αλλά με αυτόν τον τρόπο αυξάνονται επίσης τα πρότυπα τα οποία δεν κατηγοριοποιούνται σε καμία κατηγορία και χαρακτηρίζονται ως άγνωστα.

Έτσι θα πρέπει να υπάρξει απαραίτητα ένας συμβιβασμός ανάμεσα στην ακρίβεια και τον ρυθμό απόρριψης προτύπων. Ο αριθμός των προτύπων που θεωρούνται άγνωστα είναι ένας ακόμα δείκτης της σημαντικότητας των κανόνων. Ένα σύνολο κανόνων μπορεί να χαρακτηριστεί επίσης και από το μέγεθος του, όσο πιο λίγοι κανόνες υπάρχουν τόσο πιο συμπαγές και κατανοητό είναι.

6.2.3 Εξαγωγή κανόνων από όρια στον αυτο-οργανούμενο χάρτη

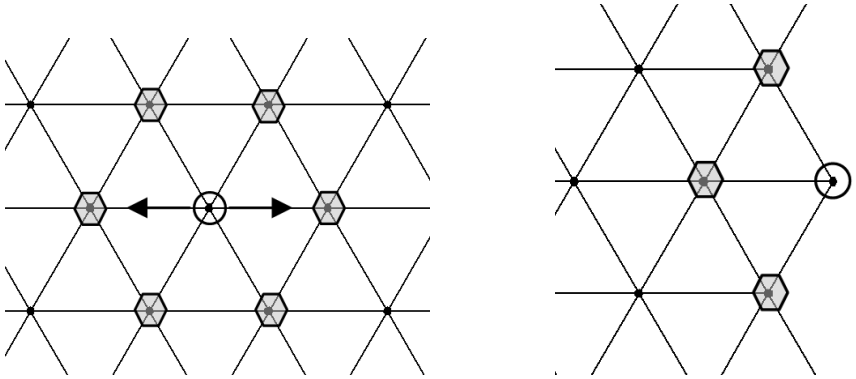
Για την δημιουργία κανόνων από αυτό-οργανούμενους χάρτες μπορεί να χρησιμοποιηθεί η ιδιότητά τους να προσφέρουν μια ομαδοποίηση των δεδομένων. Ταυτίζοντας τα όρια που σχηματίζονται μεταξύ των ομάδων στον χάρτη με όρια που σχηματίζονται για κάθε χαρακτηριστικό ξεχωριστά γίνεται η υπόθεση ότι τα όρια του χαρακτηριστικού θα ισχύουν και για τα αρχικά δεδομένα [46]. Από τα όρια των χαρακτηριστικών μπορούν να δημιουργηθούν κανόνες που αν συνδυαστούν μεταξύ τους μπορούν να δημιουργήσουν ένα σύνολο κανόνων που περιγράφει ικανοποιητικά το αρχικό σύνολο δεδομένων. Η μεθοδολογία αποτελείτε από τα εξής στάδια:

1. Δημιουργία και εκπαίδευση αυτο-οργανούμενου χάρτη
2. Υπολογισμός πίνακα που περιέχει πληροφορίες για τα όρια του χάρτη
3. Εύρεση ορίου στο χάρτη και για κάθε χαρακτηριστικό
4. Ταύτιση αρχικού ορίου με όριο στο κάθε χαρακτηριστικό
5. Εξαγωγή κανόνων όπου τα όρια στον χάρτη ταυτίζονται με τα όρια ανά χαρακτηριστικό
6. Επεξεργασία κανόνων

6.2.3.1 Εύρεση ορίων

Η εξαγωγή των κανόνων βασίζεται στην ανεύρεση ορίων που διαχωρίζουν περιοχές-ομάδες γειτονικών νευρώνων στο πλέγμα ενός εκπαιδευμένου οι οποίοι εμφανίζουν και σχετικά μικρές αποστάσεις μεταξύ τους στον χώρο των δεδομένων εισόδου. Για τον εντοπισμό την ορίων χρησιμοποιείται ο πίνακας διαφοράς ορίων (BDV - Boundary Difference Value matrix) [46]. Ο πίνακας αυτός είναι ίδιων διαστάσεων με το πλέγμα του αυτό-οργανούμενου χάρτη και κάθε τιμή του

πίνακα αντιστοιχεί και σε έναν από τους νευρώνες του πλέγματος. Για τον υπολογισμό των τιμών του ορίζεται ένα μέτρο ανομοιότητας του νευρώνα σε σχέση με τους γειτονικούς του νευρώνες προς μία όμως κατεύθυνση (σχήμα 6.9.α).



(α) (β)

Σχήμα 6.9: *(α) Πιθανή κατεύθυνση ορίου σε εσωτερικό νευρώνα του χάρτη. (β) Νευρώνας στα άκρα του χάρτη.*

Ως τιμή του πίνακα για τον κάθε νευρώνα επιλέγεται η μέγιστη τιμή του μέτρου αυτού που αντιστοιχεί σε μία από τις πιθανές κατευθύνσεις,

$$BDV = \max\left(\frac{M_L - M_O}{R_O}\right) \tag{6.32}$$

όπου M_L είναι η μέση τιμή των αποστάσεων του υπό εξέταση νευρώνα με τους νευρώνες που συμμετέχουν στο όριο, M_O είναι η μέση τιμή των αποστάσεων του υπό εξέταση νευρώνα με τους υπόλοιπους γειτονικούς νευρώνες και R_O είναι το εύρος των αποστάσεων των υπόλοιπων γειτονικών νευρώνων. Το εύρος των αποστάσεων είναι η μέγιστη απόσταση μεταξύ του υπό εξέταση νευρώνα και κάποιου γειτονικού του που δε συμμετέχει στο όριο μείον την ελάχιστη απόσταση του υπό εξέταση νευρώνα και κάποιου γειτονικού που δε συμμετέχει στο όριο.

Οι τιμές του πίνακα αυτού δηλώνουν κατά πόσο ο αντίστοιχος νευρώνας συμμετέχει σε κάποιο όριο, όσο μεγαλύτερη είναι αυτή η τιμή τόσο πιο ευδιάκριτο είναι το όριο αυτό.

Από τον παραπάνω ορισμό (εξ. 6.32) προκύπτει πρόβλημα στα άκρα του χάρτη. Όπως φαίνεται και στο σχήμα 6.9.β, οι γειτονικοί νευρώνες ενός νευρώνα που βρίσκεται στα άκρα είναι λιγότεροι. Υπάρχει η περίπτωση οι νευρώνες που συμμετέχουν στον υπολογισμό της τιμής R_O να είναι μόνο δύο. Με μόνο δυο νευρώνες αυξάνεται η πιθανότητα η απόσταση του υπό εξέταση νευρώνα από αυτούς τους δύο νευρώνες να είναι σχεδόν ίδια. Σε αυτές τις περιπτώσεις, ο παρανομαστής R_O γίνεται πολύ μικρός και η τιμή BDV πολύ μεγάλη χωρίς

όμως να δείχνει αυτό κάποιο πραγματικό όριο στο χάρτη. Κατά τον σχηματισμό των ορίων μεγάλη τιμή BDV σημαίνει σημαντικό όριο οπότε το λάθος που περιγράφηκε παραπάνω μπορεί να οδηγήσει τον αλγόριθμο στον σχηματισμό λάθος ορίων.

Ένας τρόπος για να αποφευχθεί αυτό το πρόβλημα είναι να χρησιμοποιηθεί ένας εναλλακτικός ορισμός για το BDV όπου δε θα υπάρχει το εύρος των αποστάσεων των νευρώνων, δηλαδή:

$$BDV = \max\left(M_L - M_O\right) \tag{6.33}$$

Με χρήση αυτού του τύπου σε όλο το μήκος του χάρτη υπάρχουν συγκρίσιμες τιμές BDV.

Ως όριο στον εκπαιδευμένο αυτό-οργανούμενο χάρτη ορίζεται κάθε γραμμή που αρχίζοντας από ένα νευρώνα σε κάποια άκρη του χάρτη τελειώνει σε έναν άλλο νευρώνα που βρίσκεται επίσης σε άκρη και με αυτόν τον τρόπο χωρίζει το χάρτη σε δυο μέρη. Στην παρούσα μεθοδολογία δε γίνεται αναζήτηση για κλειστά όρια που βρίσκονται αποκλειστικά στο εσωτερικό του χάρτη. Η αναζήτηση των ορίων γίνεται αρχίζοντας από τον νευρώνα που βρίσκεται σε άκρη και έχει τη μεγαλύτερη τιμή BDV. Στη συνέχεια από τους γειτονικούς νευρώνες επιλέγεται ως υποψήφιος για το όριο αυτός με τη μεγαλύτερη τιμή BDV. Επίσης γίνεται έλεγχος για το αν ο υποψήφιος νευρώνας είναι ήδη μέρος του ορίου. Αν συμβαίνει αυτό σημαίνει ότι ο σχηματισμός του ορίου πηγαίνει ξανά προς τα πίσω οπότε επιλέγεται άλλος υποψήφιος. Τέλος, ελέγχεται αν η τιμή BDV του υποψήφιου νευρώνα είναι κατά πολύ μικρότερη από τον προηγούμενο νευρώνα. Σε αυτή την περίπτωση το υπό δημιουργία όριο απορρίπτεται αφού δεν υπάρχει κάποιος γειτονικός νευρώνας με αρκετά υψηλή τιμή BDV για να συμμετάσχει στο όριο.

Αν όλες οι παραπάνω προϋποθέσεις ικανοποιούνται τότε ο υποψήφιος νευρώνας θεωρείται ότι συμμετέχει στο όριο και η αναζήτηση συνεχίζεται ψάχνοντας τους γειτονικούς του νέου νευρώνα μέχρι το όριο να φτάσει σε κάποια άκρη του χάρτη και να τον χωρίσει δυο περιοχές. Η διαδικασία συνεχίζεται επιλέγοντας τον νευρώνα που βρίσκεται σε άκρη και έχει την επόμενη μεγαλύτερη τιμή BDV και τελειώνει όταν έχουν εξεταστεί τα όρια από όλους τους νευρώνες των άκρων.

Στην αρχή της αναζήτησης, όταν εξετάζονται οι γειτονικοί νευρώνες του αρχικού νευρώνα αποκλείονται οι γειτονικοί νευρώνες που βρίσκονται επίσης στα άκρα του χάρτη. Η εύρεση ενός τέτοιου ορίου δεν έχει κάποιο νόημα αφού πρώτον όρια κατά μήκος της άκρης του χάρτη δεν ορίζουν δυο διαφορετικές περιοχές και δεύτερον εάν η αναζήτηση προχωρήσει θα οδηγηθεί στο σχηματισμό

ενός ορίου που έχει ήδη βρεθεί ή θα βρεθεί στη συνέχεια της εκτέλεσης του αλγορίθμου, αφού όλοι οι ακριανοί νευρώνες χρησιμοποιούνται ως αρχικά σημεία για την αναζήτηση ορίων.

6.2.3.2 Ταύτιση ορίων

Εφόσον έχει προκύψει κάποιο έγκυρο όριο τότε επαναλαμβάνεται η αναζήτηση για όμοιο όριο στον χάρτη χρησιμοποιώντας για τον υπολογισμό των αποστάσεων μονό τις τιμές κάθε φορά ενός χαρακτηριστικού εισόδου. Η αναζήτηση δε χρειάζεται να γίνει σε ολόκληρο τον χάρτη αλλά αρκεί να εξετασθούν οι περιοχές που βρίσκονται κοντά στους νευρώνες που αποτελούν και το αρχικό όριο. Έτσι η αναζήτηση ξεκινάει από τα δυο άκρα του αρχικού ορίου και γίνεται με παρόμοιο τρόπο με προηγουμένως.

Όταν βρεθούν κάποια έγκυρα όρια χρησιμοποιώντας μόνο τις τιμές ενός χαρακτηριστικού συγκρίνονται με το αρχικό όριο και αν κάποιο από αυτά είναι όμοιο θεωρείτε ότι το χαρακτηριστικό αυτό είναι σημαντικό για τον σχηματισμού του αρχικού ορίου και η διαδικασία συνεχίζει με την δημιουργία κανόνα αλλιώς συνεχίζει με το επόμενο χαρακτηριστικό. Η σύγκριση των δύο ορίων γίνεται με βάση το μήκος του μεγαλύτερου και εξετάζεται η ταύτιση των νευρώνων που αποτελούν τα όρια. Το ελάχιστο ποσοστό ταύτισης έτσι ώστε να θεωρηθούν δύο όρια όμοια είναι μία παράμετρος ρυθμιζόμενη από τον χρήστη. Όσο μεγαλύτερο είναι αυτό το ποσοστό τόσο λιγότερα όρια ταυτίζονται με αποτέλεσμα και λιγότερους κανόνες.

Όταν υπάρχει αυστηρότητα στην ταύτιση των ορίων επιλέγονται μόνο τα όρια που σχηματίζονται σε παρόμοιες περιοχές του χάρτη και του χάρτη της χαρακτηριστικού. Σε σύνολα δεδομένων πολλών μεταβλητών είναι δύσκολο ένα χαρακτηριστικό να είναι τόσο σημαντικό που τα όρια του να υπάρχουν αυτούσια και στο συνολικό χάρτη. Στη γενική περίπτωση τα όρια στο χάρτη οφείλονται στο συνδυασμό των χαρακτηριστικών. Έτσι, αν τα όρια πρέπει να ταυτίζονται αυστηρά μόνο ελάχιστα από αυτά θα επιλεγούν και δε θα είναι αρκετά για να δημιουργηθούν κανόνες που να περιγράφουν σωστά τα δεδομένα.

Για να επιλέγονται περισσότεροι κανόνες μπορεί να μειωθεί το ποσοστό της ταύτισης των ορίων. Όσο μειώνεται αυτό το ποσοστό τόσο περισσότερη βαρύτητα δίνεται στα όρια των χαρακτηριστικών σε βάρος των ορίων του συνολικού χάρτη. Με αυτό τον τρόπο επιλέγονται όλα τα "ευδιάκριτα" όρια στους χάρτες των χαρακτηριστικών και χρησιμοποιούνται για την παραγωγή κανόνων. Το γεγονός ότι αυτά δεν είναι πάντα σε μεγάλη αντιστοιχία με τα όρια στο χάρτη δε σημαίνει ότι δεν είναι σημαντικά αφού τα όρια των χαρακτηριστικών μπο-

ρεί να ισχύουν κατευθείαν για τα δεδομένα ή ακόμα και να συμμετέχουν στο σχηματισμό δευτερευόντων ορίων στο χάρτη που η αναζήτηση δεν μπορεί να εντοπίσει γιατί ακολουθεί άλλα πιο ευδιάκριτα όρια.

6.2.3.3 Δημιουργία κανόνων

Κάθε όριο στις τιμές ενός χαρακτηριστικού που ταυτοποιείται στον συνολικό χάρτη και στις τιμές ενός χαρακτηριστικού μπορεί να αποτελέσει τη βάση για τη δημιουργία ενός κανόνα. Ο κανόνας αυτός θα ισχύει για τις δύο περιοχές που ορίζονται στο χάρτη από το όριο. Η τιμή της συνθήκης του κανόνα μπορεί να βρεθεί από τους νευρώνες που αποτελούν το όριο. Ο μέσος όρος των τιμών του συγκεκριμένου χαρακτηριστικού όλων των νευρώνων κατά μήκος του ορίου στο χάρτη της μεταβλητής είναι η τιμή της συνθήκης στον κανόνα.

Μετά την εξαγωγή του μέσου όρου πρέπει να βρεθεί ποιες περιοχές έχουν μικρότερες τιμές από αυτή και ποιες μεγαλύτερη. Σε κάθε περιοχή του χάρτη που ορίζει το όριο λαμβάνεται ο μέσος όρος των τιμών όλων των νευρώνων. Αν αυτός ο μέσος όρος είναι μεγαλύτερος από την τιμή της συνθήκης τότε για την περιοχή αυτή ο κανόνας ισχύει για τιμές μεγαλύτερες της συνθήκης αλλιώς για μικρότερες. Η χρήση του μέσου όρου σε μεγάλη περιοχή του χάρτη και μεγάλο αριθμό νευρώνων κρύβει κινδύνους γιατί μικρές ομάδες νευρώνων μπορεί να έχουν πολύ διαφορετικές τιμές από την υπόλοιπη περιοχή και έτσι ο κανόνας να μην ισχύει για αυτές. Τέτοιες περιπτώσεις μπορούν να ανακαλυφθούν και να διορθωθούν στο στάδιο επεξεργασίας των κανόνων.

Στη συνέχεια πρέπει να γίνει αντιστοίχηση των δύο περιοχών που ορίστηκαν στο χάρτη με τις κατηγορίες των δεδομένων εισόδου. Κάθε νευρώνας κατηγοριοποιείται με βάση των πλειοψηφούσα κατηγορία των προτύπων που αντιστοιχούν σε αυτόν μετά την εκπαίδευση του χάρτη (BMUs). Η δημιουργία των κανόνων για μία από τις δύο περιοχές πραγματοποιείται στις εξής περιπτώσεις:

1. Οι νευρώνες μιας κατηγορίας να αποτελούν πλειοψηφία σε μια περιοχή του χάρτη. Ακόμα και αν οι νευρώνες αυτοί είναι πολύ λιγότεροι από τους νευρώνες της κατηγορίας στο σύνολο του χάρτη μπορούν να δημιουργηθούν κανόνες που να περιγράφουν καλά ένα μέρος των δεδομένων αυτής της κατηγορίας αποκλείοντας παράλληλα δεδομένα άλλων κατηγοριών.

2. Οι νευρώνες μίας κατηγορίας να βρίσκονται ως επί το πλείστον στην υπό εξέταση περιοχή του χάρτη. Δηλαδή, απαιτείται το πλήθος των νευρώνες της κατηγορίας στην περιοχή να υπερβαίνει ένα προκαθορισμένο ποσοστό του συνόλου των νευρώνων της κατηγορίας. Σε αυτήν την περίπτωση

> δημιουργείτε ένας κανόνας για την άλλη περιοχή, ο οποίος όμως είναι αρνητικός ως προς την ύπαρξη της κατηγορίας στην περιοχή αυτή

Το αποτέλεσμα αυτής της διαδικασίας είναι η δημιουργία μεγάλου αριθμού κανόνων για κάθε κατηγορία. Οι κανόνες αυτοί έχουν από ένα κατηγόρημα. Για την απαλοιφή λανθασμένων κανόνων και για το συνδυασμό των καλύτερων κανόνων κάθε κατηγορίας είναι απαραίτητο ένα επιπλέον στάδιο επεξεργασίας κανόνων.

6.2.3.4 Επεξεργασία κανόνων

Οι κανόνες που κατασκευάζονται με την παραπάνω διαδικασία σε λίγες μόνο περιπτώσεις καταφέρνουν να περιγράψουν με ακρίβεια και αποκλειστικά μία κατηγορία δεδομένων. Πιο αποδοτική περιγραφή μπορεί να γίνει με συνδυασμό των κανόνων. Ο συνδυασμός των κανόνων μπορεί να γίνει με δύο τρόπους, με σύζευξη (AND) ή διάζευξη (OR). Βρέθηκε ότι οι περισσότεροι κανόνες εκτός από την κατηγορία που περιγράφουν επαληθεύονται και από αρκετά άλλα πρότυπα. Για το λόγο αυτό γίνεται συνδυασμός κανόνων κυρίως με σύζευξη. Κανόνες που δημιουργήθηκαν στην περίπτωση που οι νευρώνες μιας κατηγορίας ήταν η πλειοψηφία σε μια περιοχή του χάρτη μπορούν να συνδυαστούν με διάζευξη αν επαληθεύονται από λίγα λάθος πρότυπα.

Κάθε κανόνας εφαρμόζεται στο αρχικό σύνολο δεδομένων για να καταμετρηθούν πόσα και ποιας κατηγορίας πρότυπα τον επαληθεύουν. Αν ένας κανόνας δεν περιγράφει σωστά το 50% των προτύπων της κατηγορίας η συνθήκη του αντιστρέφεται και ισχύει πλέον ο νέος κανόνας. Αυτό γίνεται ώστε να διορθωθούν τυχόν λάθη που οφείλονται στη χρήση μέσου όρου για την κατασκευή του κανόνα.

Στην συνέχεια χρησιμοποιούνται οι ποσότητες n_{++} και n_{+-} αφού όπως αναφέρθηκε αρκούν για να υπολογιστούν τα μέτρα αποδοτικότητας. Το μέτρο που χρησιμοποιήθηκε είναι το S_2.

Ανάμεσα στους απλούς κανόνες υπάρχουν κάποιοι που περιγράφουν μια κατηγορία πλήρως. Οι κανόνες αυτοί συνδέονται με AND και σχηματίζουν τον κανόνα της κατηγορίας αυτής. Αν ο κανόνας που προκύπτει δεν είναι ικανοποιητικός τότε πρέπει να εφαρμοστούν και άλλοι κανόνες. Οι υπόλοιποι κανόνες μπορούν να συνδυαστούν με δύο διαφορετικούς τρόπους που καταλήγουν σε διαφορετικά σύνολα κανόνων.

Στον πρώτο τρόπο οι κανόνες ταξινομούνται με βάση την ποσότητα n_{++} και αν υπάρχει ισότητα και με βάση την n_{+-}. Οι κανόνες με καλύτερη περιγραφή της κατηγορίας, δηλαδή μεγάλο n_{++} και μικρό n_{+-} συνδυάζονται ένας κάθε

φορά στον κανόνα της κατηγορίας που έχει κατασκευαστεί μέχρι εκείνη τη στιγμή μέχρι ο κανόνας να μην κατηγοριοποιεί καθόλου λάθος πρότυπα οπότε και δεν υπάρχει η ανάγκη προσθήκης άλλων κανόνων. Αν η εφαρμογή ενός κανόνα χειροτερεύει τα αποτελέσματα που υπήρχαν μέχρι εκείνη τη στιγμή τότε ο κανόνας αυτός απορρίπτεται. Στο δεύτερο τρόπο οι κανόνες ταξινομούνται με βάση το κριτήριο S_2 και συνδυάζονται όπως και στην πρώτη περίπτωση.

Με αυτή τη διαδικασία κατασκευάζονται δυο διαφορετικά σύνολα κανόνων που έχουν έναν κανόνα για κάθε κατηγορία του αρχικού συνόλου δεδομένων. Οι κανόνες αυτοί αποτελούνται από ένα ή περισσότερα κατηγορήματα το πλήθος των οποίων είναι πάντα κατά πολύ μικρότερο από το πλήθος των κανόνων που είχαν κατασκευαστεί αρχικά. Για κάθε κατηγορία επιλέγεται ο ένας από τους δύο κανόνες που την περιγράφει πιο καλά και έτσι δημιουργείται το τελικό σύνολο κανόνων.

6.2.4 Εξαγωγή κανόνων από ομάδες στον αυτο-οργανούμενο χάρτη

Η δεύτερη προσέγγιση για την εξαγωγή κανόνων από έναν αυτο-οργανούμενο χάρτη βασίζεται στην επεξεργασία ομάδων νευρώνων που σχηματίζονται στον εκπαιδευμένο χάρτη. Η προσέγγιση αυτή είναι παρόμοια με την προηγούμενη με την διαφορά ότι οι περιοχές του χάρτη για τις οποίες εξάγονται οι κανόνες προκύπτουν από ομαδοποίηση των νευρώνων και όχι από την αντίστροφη διαδικασία που είναι η διχοτόμηση του χάρτη μέσω των ορίων. Τα σταδία αυτής της διαδικασίας είναι τα εξής:

1. Δημιουργία και εκπαίδευση αυτο-οργανούμενου χάρτη
2. Ομαδοποίηση νευρώνων του χάρτη
3. Υπολογισμός σημαντικών χαρακτηριστικών για την κάθε ομάδα
4. Εξαγωγή κανόνων για κάθε ομάδα στον χάρτη ανά σημαντικό χαρακτηριστικό
5. Επεξεργασία κανόνων

Το δεύτερο στάδιο της μεθόδου υλοποιήθηκε με χρήση της μεθοδολογίας που περιγράφηκε στην παράγραφο 3.2, ενώ για το τρίτο βήμα χρησιμοποιήθηκε ο αλγόριθμος sig* [78].

6.2.4.1 Επιλογή σημαντικών χαρακτηριστικών

Για την επιλογή των σημαντικών χαρακτηριστικών ανά ομάδα χρησιμοποιήθηκε ο αλγόριθμος sig*. Ο σχετικά απλός αυτός αλγόριθμός επιλεγεί τα σημαντικά χαρακτηριστικά για κάθε ομάδα εφόσον παρέχεται ένας βαθμός αξιολόγησης των χαρακτηριστικών. Στην συγκεκριμένη περίπτωση για την εξαγωγή του βαθμού αξιολόγησης χρησιμοποιήθηκαν δυο προσεγγίσεις. Η πρώτη είναι η χρήση του λόγου της τυπικής απόκλισης των τιμών του κάθε χαρακτηριστικού σε όλο το χάρτη προς την τυπική απόκλιση των τιμών του χαρακτηριστικού στην ομάδα. Η προσέγγιση αυτή βασίζεται στην θεώρηση ότι όταν οι τιμές ενός χαρακτηριστικών παρουσιάζουν μειωμένη μεταβλητότητα σε κάποια ομάδα τότε μπορεί να θεωρηθεί ότι συνεισφέρει σημαντικά στον σχηματισμό της ομάδας. Η προσέγγιση αυτή είναι σχετικά απλοϊκή και χρησιμοποιήθηκε κυρίως για λόγους σύγκρισης. Η δεύτερη προσέγγιση για την αξιολόγηση των χαρακτηριστικών βασίζεται στην μεθοδολογία που παρουσιάστηκε στην παράγραφο 4.2.1.

Ο αλγόριθμος SIG*

Ο αλγόριθμος sig* για την επιλογή των χαρακτηριστικών που χαρακτηρίζουν μια ομάδα χρησιμοποιεί έναν βαθμό αξιολόγησης για κάθε χαρακτηριστικό. Για να εξηγηθεί ο αλγόριθμος έστω ένα σύνολο δεδομένων 4 χαρακτηριστικών και από το οποίο προκύπτουν 6 ομάδες. Χρησιμοποιώντας έναν βαθμό αξιολόγησης των χαρακτηριστικών κατασκευάζεται ο πίνακας 6.3.

Πίνακας 6.3: *Πίνακας βαθμών αξιολόγησης χαρακτηριστικών.*

	Ομάδες					
Χαρακτηριστικά	1	2	3	4	5	6
1	4.535	2.673	4.2241	2.5619	3.3118	5.4126 *
2	2.0027	2.0065	3.4938 *	2.1526	2.4081	2.0771
3	10.004	12.986 *	5.1911	5.4576	3.479	7.1688
4	10.21	13.626 *	3.0457	4.7029	4.4746	6.3346

Στον πίνακα 6.3 η μεγαλύτερη τιμή κάθε χαρακτηριστικού μαρκάρεται με ένα αστέρι (*). Για την επιλογή των πιο σημαντικών χαρακτηριστικών για την περιγραφή κάθε ομάδας σχηματίζεται έναν νέος πίνακας όπου οι τιμές σπουδαιότητας κάθε χαρακτηριστικού κανονικοποιούνται ως προς το άθροισμα των τιμών σπουδαιότητας για όλη την ομάδα. Οι τιμές αυτές ταξινομούνται από την μεγαλύτερη προς τη μικρότερη και υπολογίζονται οι αθροιστικοί βαθμοί αξιολόγησης. Για παράδειγμα για τις ομάδες 1 και 6 οι νέες τιμές παρουσιάζονται

στον πίνακα 6.4.

Πίνακας 6.4: *Ιεράρχηση χαρακτηριστικών για τις ομάδες 1 και 6.*

	Χαρακτηριστικά	Βαθ. Αξιολόγησης (%)	Αθροιστικός Βαθ. Αξιολόγησης (%)
Ομάδα 1	4	38.1665	38.1665
	3	37.3952	75.5617
	1	16.9521	92.5138
	2	7.4862	100
Ομάδα 6	3	34.1484	34.1484
	4	30.1748	64.3231
	1 *	25.7828	90.1060
	2	9.8940	100

Από τον πίνακα 6.4 επιλέγονται τα χαρακτηριστικά με το μεγαλύτερο ποσοστό αξιολόγησης μέχρι οι αθροιστικές τιμές να είναι ίσες ή να ξεπεράσουν ένα προκαθορισμένο κατώφλι. Μια συνηθισμένη επιλογή για το κατώφλι αυτό είναι το 50%. Με αυτό το κατώφλι στην ομάδα 1 επιλέγονται τα χαρακτηριστικά 4 και 3 ενώ στην ομάδα 6 τα χαρακτηριστικά 3 και 4. Επειδή στην ομάδα 6 το χαρακτηριστικό 1 παίρνει τη μεγαλύτερη τιμή σημαντικότητας και είναι σημαδεμένο με αστεράκι πρέπει να θεωρηθεί ότι είναι σημαντικό για την ομάδα ανεξάρτητα από το γεγονός ότι η αθροιστική τιμή σπουδαιότητας υπερβαίνει το κατώφλι. Γενικά, όλα τα χαρακτηριστικά που έχουν αστεράκι θεωρούνται σημαντικά για την αντίστοιχη ομάδα.

Από τη στιγμή που έχουν επιλεγεί τα χαρακτηριστικά που περιγράφουν ικανοποιητικά κάθε ομάδα μπορούν να χρησιμοποιηθούν για την κατασκευή κανόνων. Αν οι συνθήκες των κανόνων είναι πολύ αυστηρές τότε πολλά πρότυπα που ανήκουν στην ομάδα μπορεί να ικανοποιούν τους κανόνες αυτούς. Αντίθετα, αν οι συνθήκες των κανόνων είναι πολύ χαλαρές τότε θα τις ικανοποιούν πρότυπα που κανονικά δεν ανήκουν σε αυτή την ομάδα οδηγώντας και πάλι σε λανθασμένη κατηγοριοποίηση. Για να κατασκευαστούν σωστοί κανόνες πρέπει να γίνει μια καλή εκτίμηση για την κατανομή των δεδομένων μέσα στην ομάδα.

6.2.4.2 Εξαγωγή και επεξεργασία κανόνων

Οι πιο απλοί κανόνες κατασκευάζονται υπολογίζοντας την ελάχιστη και την μέγιστη τιμή του κάθε χαρακτηριστικού σε κάθε ομάδα και θεωρώντας αυτές τις

τιμές ως κάτω και άνω όριο του κανόνα αντίστοιχα. Όμως με τέτοιου τύπου κανόνες δε γίνεται κάποια υπόθεση για την κατανομή των δεδομένων και μπορεί να κατασκευαστούν κανόνες μικρής ακρίβειας αφού αρκεί ένα μόνο πρότυπο που να έχει τιμή αρκετά μικρότερη ή αρκετά μεγαλύτερη από όλα τα υπόλοιπα της ομάδας για να αλλάξει τελείως η συνθήκη του κανόνα με αποτέλεσμα χειρότερη κατηγοριοποίηση.

Μια υπόθεση που γίνεται συχνά και εφαρμόστηκε και σε αυτήν την περίπτωση είναι ότι τα δεδομένα ακολουθούν την κανονική κατανομή. Μια σημαντική ιδιότητα της κανονικής κατανομής που είναι γνωστή από την στατιστική είναι ότι το 95% των δεδομένων περιέχονται στο διάστημα $\mu \pm 2\sigma$ όπου μ είναι η μέση τιμή των δεδομένων και σ η τυπική απόκλιση των δεδομένων. Έτσι για κάθε σημαντικό χαρακτηριστικό j κατασκευάζονται κανόνες με όρια τις τιμές $\mu_j - 2\sigma_j$ και $\mu_j + 2\sigma_j$. Η μέση τιμή και η τυπική απόκλιση κάθε ομάδας υπολογίζονται από τα αποκανονικοποιημένα βάρη των νευρώνων έτσι ώστε οι κανόνες να ισχύουν απευθείας για το αρχικό σύνολο δεδομένων και να έχουν μεγαλύτερη φυσική σημασία. Η περαιτέρω επεξεργασία των κανόνων γίνεται με τον ίδιο τρόπο που εφαρμόστηκε και στην προηγούμενη μεθοδολογία.

6.2.5 Πειραματική αξιολόγηση

Με στόχο την αξιολόγηση των εξαγόμενων κανόνων ως προς την ικανότητα ορθής κατηγοριοποίησης των δεδομένων, πραγματοποιήθηκε πειραματική αξιολόγηση με την χρήση τριών συνόλων δεδομένων από το UCI machine-learning repository [8]. Τα σύνολα αυτά είναι τα σύνολα Ionosphere, Iris και Image Segmentation, τα οποία περιγράφονται στο Παράρτημα Α. Στα πειράματα δοκιμάστηκαν τέσσερις διαφορετικές προσεγγίσεις. Οι δύο πρώτες υλοποιούν την μεθοδολογία εξαγωγής μέσω κανόνων με την χρήσης των ορίων με την πρώτη να υλοποιεί την μέθοδο με χρήση της εξίσωσης 6.32, ενώ η δεύτερη χρησιμοποιεί την εξίσωση 6.33. Η τρίτη και η τέταρτη βασίζονται στην εξαγωγή κανόνων από ομάδες και χρησιμοποιούν ως βαθμό αξιολόγησης τον λόγο των τυπικών αποκλίσεων και την μέθοδο της παραγράφου 4.2.1 αντίστοιχα.

Η εκπαίδευση των αυτο-οργανούμενων χαρτών και όλη η διαδικασία εξαγωγής των κανόνων καθώς και η αξιολόγηση έγινε με όλα τα πρότυπα των συνόλων. Ο στόχος ήταν να διερευνηθεί η δυνατότητα των μεθόδων να εξάγουν κανόνες που περιγραφούν ικανοποιητικά το σύνολο των δεδομένων και όχι η ικανότητα γενίκευσης των κανόνων σε νέα δεδομένα. Τα αποτελέσματα παρουσιάζονται στον πίνακα 6.5

Η ρύθμιση των παραμέτρων του αυτο-οργανούμενο χάρτη γίνεται με τον

Πίνακας 6.5: *Ποσοστά ορθής κατηγοριοποίησης με χρήση των εξαγόμενων κανόνων.*

		Απόδοση (%)		
Μέθοδος	Μέγεθος χάρτη	Ionosphere	Image segm.	Iris
1	Κανονικό	92.59	88.48	96.67
1	Μεγάλο	94.87	91.73	-
2	Κανονικό	90.88	86.45	96.67
2	Μεγάλο	94.87	91.52	-
3	Κανονικό	90.31	77.23	95.33
3	Μεγάλο	90.60	76.93	-
4	Κανονικό	95.73	94.16	97.33
4	Μεγάλο	95.73	94.37	-

τρόπο που περιγράφεται στην παράγραφο 5.4.1. Για το σύνολο δεδομένων Iris χρησιμοποιήθηκε μόνο ένα μέγεθος χάρτη καθώς το πλήθος των δεδομένων δεν επέτρεπε την χρήση μεγαλύτερου μεγέθους χάρτη.

Ο αυτο-οργανούμενος χάρτης εκπαιδεύτηκε από τα δεδομένα χωρίς γνώση της κατηγορίας των προτύπων, η οποία δεν χρησιμοποιήθηκε ούτε στο στάδιο της ανεύρεσης των ορίων ή των ομάδων. Η γνώση αυτή χρησιμοποιήθηκε μόνο κατά το στάδιο της εξαγωγής των κανόνων και της επεξεργασίας των κανόνων.

Οι δύο πρώτες μέθοδοι παρουσίασαν καλή απόδοση και στις τρεις περιπτώσεις και όπως φαίνεται και από τα αποτελέσματα η απόδοση αυτή επηρεάζεται από το μέγεθος του χάρτη. Η τρίτη μέθοδος παρουσίασε τα χειρότερα αποτελέσματα και αυτό γιατί ο βαθμός αξιολόγησης που χρησιμοποιήθηκε δεν είναι ιδιαίτερα περιγραφικός για την σημαντικότητα του κάθε χαρακτηριστικού στον σχηματισμό των ομάδων. Η τέταρτη μέθοδος παρουσίασε τα καλύτερα αποτελέσματα και στις τρεις περιπτώσεις και δεν φαίνεται να επηρεάζεται από το μέγεθος του χάρτη.

6.3 Συζήτηση - Συμπεράσματα

Στην πρώτη ενότητα αυτού του κεφαλαίου παρουσιάστηκε ένα υβριδικό σύστημα αναγνώρισης χειρονομιών. Το σύστημα αυτό εντάσσεται στην κατηγορία των υβριδικών συστημάτων μετασχηματισμού, διότι η κύρια λειτουργία του πρώτου τμήματος του συστήματος είναι να μετασχηματίσει την πληροφορία σε συμβολική μορφή έτσι ώστε να είναι δυνατή η δημιουργία των πιθανοτικών μοντέλων.

Ο μετασχηματισμός αυτός έγινε με χρήση ενός αυτο-οργανούμενου χάρτη, του οποίου οι δυνατότητες κρίθηκαν ως κατάλληλες για να διεκπεραιώσουν αυτήν λειτουργία. Επίσης, το σύστημα παρουσιάζει αμφίδρομη ροή πληροφοριών καθώς κατά την διαδικασία κατηγοριοποίησης ενός νέου στιγμιότυπου χειρονομίας, η οποία γίνεται με την χρήση των μοντέλων Markov, τα μοντέλα αυτά χρησιμοποιούν τον αυτο-οργανούμενο χάρτη για να αντλήσουν την πληροφορία την γειτνίασης μεταξύ των πιθανών καταστάσεων μετάβασης.

Παρόλο που το σύστημα αυτό βρίσκεται ακόμα στο αρχικό στάδιο ανάπτυξής του, τα πρώτα πειραματικά αποτελέσματα είναι άκρως ενθαρρυντικά για την περαιτέρω πορεία της έρευνας. Ενδιαφέρον επίσης έχει και η εφαρμογή της προσέγγισης αυτής σε αλλά προβλήματα εκτός αυτού της αναγνώρισης των χειρονομιών, τα οποία παρουσιάζουν παρόμοια χαρακτηριστικά.

Στην δεύτερη ενότητα παρουσιάστηκαν επίσης μεθοδολογίες μετασχηματισμού. Στην περίπτωση αυτή, ο κύριος σκοπός αυτών των μεθοδολογιών δεν είναι να δημιουργήσουν ένα σύνολο κανόνων που θα λειτουργήσει ως σύστημα κατηγοριοποίησης, αλλά να εξάγουν γνώση σε μορφή πιο χρηστική και κατανοητή για τον άνθρωπο. Οι αυτο-οργανούμενοι χάρτες είναι ένα μοντέλο που παρέχει δυνατότητες οπτικοποίησης των δεδομένων. Αυτές οι δυνατότητες σε συνδυασμό με ένα σύνολο κανόνων, προσφέρουν στον αναλυτή ευκολότερη κατανόηση των δεδομένων, γεγονός πολύ χρήσιμο σε εφαρμογές εξόρυξης γνώσης σε πραγματικά περιβάλλοντα. Οι μεθοδολογίες που παρουσιάστηκαν κατορθώνουν να δημιουργήσουν σύνολα κανόνων τα οποία περιγράφουν ικανοποιητικά τα δεδομένα και τις κατηγορίες που αυτά ανήκουν παρόλο που η εκπαίδευση του συστήματος γίνεται χωρίς επίβλεψη, δηλαδή το σύστημα στο πρώτο στάδιο δεν έχει επίγνωση των κατηγοριών των δεδομένων.

□

Κεφάλαιο 7

Συνολικό πόρισμα διατριβής

7.1 Γενικά συμπεράσματα

Κατά την εκπόνηση της παρούσας διατριβής, πραγματοποιήθηκε η ερευνητική δραστηριότητα σε διάφορα πεδία, όπως προκύπτει και από το παρόν σύγγραμμα. Το γεγονός αυτό συνάδει με το αρχικό στόχο της διατριβής ο οποίος ήταν ο σχεδιασμός, η υλοποίηση και η αξιολόγηση υβριδικών τεχνικών. Ξεκινώντας από τις μεθοδολογίες ομαδοποίησης που παρουσιάστηκαν στο τρίτο κεφάλαιο ως τις προσεγγίσεις στο ζήτημα της εξαγωγής συμβολικής γνώσης που περιγράφονται στο έκτο κεφάλαιο, όλες οι μεθοδολογίες και τα συστήματα παρουσιάζουν κάποιο μικρό ή μεγάλο βαθμό υβριδικότητας και μπορούν να ενταχθούν σε μία περισσότερες από τις κατηγορίες υβριδικών συστημάτων που καταγράφονται στο δεύτερο κεφάλαιο.

Ένα από τα γενικά συμπεράσματα που προκύπτει από την ερευνητική δραστηριότητα αυτής της διατριβής είναι ότι τα υβριδικά συστήματα αποτελούν ένα σημαντικό κεφάλαιο στην μελλοντική ανάπτυξη των πεδίων της υπολογιστικής νοημοσύνης και της μηχανικής μάθησης. Η ως τώρα ερευνητική δραστηριότητα έχει καταδείξει και καταγράψει τα μειονεκτήματα και τις αδυναμίες των υπαρχόντων μεθοδολογιών και συστημάτων, οι οποίες είναι αρκετές φόρες δομικές και συνεπώς σχεδόν αδύνατον να ξεπεραστούν. Τα υβριδικά συστήματα προσφέρουν μία διέξοδο από αυτήν κατάσταση καθώς όπως έδειξε και η παρούσα διατριβή παρέχουν την δυνατότητα βελτίωσης των μεμονωμένων συστημάτων και κάλυψη των αδυναμιών τους.

Σε όλες τις μεθοδολογίες και τα συστήματα που αναπτύχθηκαν ένα από τα συνιστώσα τμήματα είναι το μοντέλο των αυτο-οργανούμενων χαρτών (SOM). Το μοντέλο αυτό παρά την φαινομενική απλότητά του, προσφέρει πλήθος δυνατοτήτων και μπορεί να αποτελέσει τμήμα ενός υβριδικού συστήματος εξυπηρετώ-

ντας κάθε φορά διαφορετικούς σκοπούς. Το γεγονός αυτό ενισχύεται και από την εντεινόμενη ερευνητική δραστηριότητα στο πεδίο αυτό που αποδεικνύεται από την αυξανόμενη σχετική βιβλιογραφία.

Το ίδιο συμπέρασμα προκύπτει αβίαστα και από την παρούσα διατριβή καθώς είναι εμφανής η πολύπλευρη αξιοποίηση του μοντέλου στις μεθοδολογίες που αναπτύχθηκαν. Πιο συγκεκριμένα, στην μεθοδολογία ομαδοποίησης που παρουσιάστηκε στο τρίτο κεφάλαιο, οι αυτο-οργανούμενοι χάρτες χρησιμοποιήθηκαν ως αρχικό στάδιο ομαδοποίησης των δεδομένων. Στο τέταρτο κεφάλαιο αναπτύχθηκε μία μέθοδος αξιολόγησης των χαρακτηριστικών των δεδομένων εισόδου ως προς την συνεισφορά τους στον σχηματισμό των ομάδων που σχηματίστηκαν από τα δεδομένα εισόδου. Σε αυτήν την περίπτωση έγινε εκμετάλλευση της δυνατότητας επεξεργασίας μεμονωμένων χαρακτηριστικών εισόδου διατηρώντας όμως την πληροφορία για την τοπολογική συσχέτιση των προτύπων.

Στο σύστημα κατηγοριοποίησης που περιγράφηκε στο πέμπτο κεφάλαιο, οι αυτο-οργανούμενοι χάρτες χρησιμοποιήθηκαν ως εκτιμητές πιθανότητας για διαφορετικούς συνδυασμούς χαρακτηριστικών που προκύπτουν κατά την διαδικασία κατηγοριοποίησης. Τέλος, στο έκτο κεφάλαιο υλοποιήθηκαν μεθοδολογίες στις οποίες οι αυτο-οργανούμενοι χάρτες χρησιμοποιήθηκαν στον μετασχηματισμό πληροφοριών σε συμβολική μορφή.

7.2 Μελλοντικές επεκτάσεις

Από την παρούσα διατριβή προκύπτουν διάφορα θέματα που αποτελούν πεδίο μελλοντικής έρευνας με στόχο είτε την αντιμετώπιση των αδυναμιών των μεθοδολογιών που παρουσιάστηκαν στην διατριβή, είτε την δημιουργία νέων μεθόδων και προσεγγίσεων.

Οι δύο μεθοδολογίες ομαδοποίησης ακολουθούν την ιεραρχική συγκεντρωτική προσέγγιση κατά την φάση του σχηματισμού των τελικών ομάδων. Το κύριο πρόβλημα αυτής της προσέγγισης είναι ο κατάλληλος τερματισμός της διαδικασίας των συγχωνεύσεων των ομάδων. Οι δύο μεθοδολογίες χρησιμοποιούν μέτρα αξιολόγησης έτσι ώστε να τερματίσουν την διαδικασία αυτής. Τα μέτρα αυτά απαιτούν την αξιολόγησή τους από ανθρώπινο παράγοντα. Θα ήταν ενδιαφέρον να αναζητηθούν άλλοι τρόποι τερματισμού της διαδικασίας συγχώνευσης που να μην απαιτούν τέτοιου είδους αξιολόγηση.

Επίσης, ενδιαφέρον θα είχε ο συνδυασμός των δύο μεθοδολογιών. Η πρώτη μεθοδολογία χρησιμοποιεί το βαθμό συμμετοχής του κάθε προτύπου στην κάθε ομάδα, δηλαδή το κάθε πρότυπο ανήκει σε πολλές ομάδες ανάλογα με το βαθμό συμμετοχής και όχι αποκλειστικά μόνο σε μία. Η δεύτερή μεθοδολογία είναι

δυνατόν να ενσωματώσει αυτή την προσέγγιση κάνοντας χρήση ασαφών αυτο-οργανούμενων χαρτών.

Η μεθοδολογία αξιολόγησης των χαρακτηριστικών που χρησιμοποιήθηκε στον νευρο-ασαφή κατηγοριοποιητή μπορεί να συνδυαστεί και με άλλα συστήματα κατηγοριοποίησης, τα οποία μπορούν να αξιοποιήσουν τα μετα-δεδομένα που παράγονται. Με αυτόν τον τρόπο θα τεκμηριωνόταν ισχυρότερα η ικανότητα της μεθόδου στην αξιολόγηση των χαρακτηριστικών. Αξίζει, επίσης να διερευνηθεί η δυνατότητα εξαγωγής και άλλων μετα-δεδομένων με στόχο την καλύτερη παραμετροποίηση του νευρο-ασαφή κατηγοριοποιητή ή παρόμοιων συστημάτων που μπορούν να χρησιμοποιηθούν στην θέση του.

Στην περίπτωση του κατηγοριοποιητή που βασίζεται στο μοντέλο των κ - πλησιέστερων γειτόνων, πρόκληση αποτελεί η θεωρητική απόδειξη της ικανότητας αύξησης των ποσοστών κατηγοριοποίησης συγκριτικά με το μοντέλο των κ - πλησιέστερων γειτόνων και τις παραλλαγές του.

Μια διαδικασία μάθησης για την ρύθμιση των παραμέτρων k και Confidence, αποτελεί επίσης μία ενδιαφέρουσα ερευνητική προέκταση, η οποία θα έδινε και την δυνατότητα της ρύθμισης των παραμέτρων αυτών με διαφορετικές τιμές για κάθε χαρακτηριστικό των δεδομένων εισόδου. Ο κατηγοριοποιητής αυτός δεν εκμεταλλεύεται όλες τις δυνατότητες του μοντέλου των αυτο-οργανούμενων χαρτών. Οι δυνατότητες οπτικοποίησης που παρέχει το μοντέλο μπορούν να αξιοποιηθούν για να παράγεται μία οπτική αναπαράσταση των ομάδων των συνδυασμών των χαρακτηριστικών που χρησιμοποιούνται στην διαδικασία της κατηγοριοποίησης έτσι ώστε να είναι δυνατή η μελέτη και αξιολόγησή τους από τον άνθρωπο, είτε για να παρέχεται πιθανή αιτιολόγηση των αποφάσεων του κατηγοριοποιητή, είτε για καλύτερή κατανόηση των δεδομένων εισόδου και του προβλήματος που απεικονίζουν.

Μία ακόμα παραλλαγή του κατηγοριοποιητή που μπορεί να αποτελέσει αντικείμενο μελλοντικής έρευνας είναι η αντικατάσταση των αυτο-οργανούμενων χαρτών με ένα μοντέλο επιβλεπόμενης μάθησης, καθώς το σύνολο των δεδομένων με το οποίο εκπαιδεύεται ο αυτο-οργανούμενος χάρτης είναι κατηγοριοποιημένο και επομένως είναι δυνατόν να χρησιμοποιηθεί για την εκπαίδευση ενός μοντέλου επιβλεπόμενης μάθησης.

Ένα ακόμη θέμα που παρουσιάζει ερευνητικό ενδιαφέρον είναι η εφαρμογή της φιλοσοφίας του συστήματος αναγνώρισης χειρονομιών σε άλλα προβλήματα, τα οποία παρουσιάζουν παρόμοια χαρακτηριστικά, δηλαδή δεδομένα εισόδου που σχηματίζουν ακολουθίες, τέτοια προβλήματα είναι η κατηγοριοποίηση κειμένου, η εξόρυξη γνώσης στο διαδίκτυο, πρόβλεψη τροχιάς κ.α. Στο πεδίο της εξαγωγής κανόνων από τα δεδομένα, ενδιαφέρον παρουσιάζει η υλοποίηση μία με-

θοδολογίας εξαγωγής ασαφών κανόνων καθώς και η χρήση παραλλαγών του μοντέλου των αυτο-οργανούμενων χαρτών που χρησιμοποιούν χάρτες πολλαπλών επιπέδων για την καλύτερη περιγραφή των δεδομένων.

Παράρτημα Α

Χαρακτηριστικά Προβλήματα Ταξινόμησης

Α1. Σύνολα Δεδομένων UCI

Η βάση δεδομένων UCI [8] αποτελεί ίσως τη μεγαλύτερη αποθήκη πειραματικών δεδομένων και οι περισσότερες ερευνητικές μελέτες στη βιβλιογραφία απευθύνονται εκεί για την επιλογή του κατάλληλου υλικού για τις δοκιμές τους και την αξιολόγηση των μεθόδων τους. Στην παρούσα διατριβή χρησιμοποιήθηκαν τα παρακάτω σύνολα δεδομένων, τα οποία βρίσκονται διαθέσιμα στον ιστοχώρο: http://www.ics.uci.edu/m̃learn/MLRepository.html.

Σύνολο δεδομένων Breast Cancer Wisconsin

Η συγκεκριμένη βάση δεδομένων για καρκίνο του μαστού προέρχεται από το πανεπιστήμιο του Wisconsin των Η.Π.Α. και περιλαμβάνει την ταξινόμηση όγκων σε καλοήθεις και κακοήθεις, βάσει 9 χαρακτηριστικών που περιγράφουν τη φυσιολογία των κυττάρων. Το σύνολο δεδομένων περιλαμβάνει 699 πρότυπα.

Σύνολο δεδομένων Covertype

Το σύνολο Covertype (The Forest CoverType dataset) προέρχεται από το τμήμα δασικών επιστημών του Colorado State University. Κάθε πρότυπο αποτελείτε από εδαφολογικές και τοπογραφικές μετρήσεις σε ένα τετράγωνο διαστάσεων 30x30 μέτρων. Οι μετρήσεις αυτές αφορούν τετράγωνα δασικών εκτάσεων και ο σκοπός είναι ο προσδιορισμός του είδους του δένδρου που καλύπτει την περιοχή. Τα δεδομένα είναι ταξινομημένα σε 7 κατηγορίες που κάθε μία αντιστοιχεί και σε διαφορετικό είδος δένδρου. Ο αριθμός των προτύπων είναι ιδιαί-

τερα μεγάλος καθώς το σύνολο αποτελείτε από 581012 πρότυπα. Κάθε πρότυπο χαρακτηρίζεται από 54 τιμές.

Σύνολο δεδομένων Glass

Το σύνολο δεδομένων Glass (Glass Identification Database) περιλαμβάνει δεδομένα που περιγραφούν διαφόρους τύπους γυαλιού. Τα 9 χαρακτηριστικά του κάθε προτύπου είναι τα διάφορα χημικά στοιχεία από τα οποία αποτελείτε το συγκεκριμένο δείγμα γυαλιού. Τα 214 πρότυπα είναι κατηγοριοποιημένα σε 7 διαφορετικές κατηγορίες που αφορούν την χρήση του συγκεκριμένου είδους γυαλιού, π.χ υαλοπίνακες κτηρίων, οχημάτων κλπ.

Σύνολο δεδομένων Image Segmentation

Το σύνολο δεδομένων Image Segmentation αποτελείται από 2310 πρότυπα με 19 γνωρίσματα, τα οποία περιγράφουν μέρος μίας εικόνας. Το σύνολο έχει προέλθει από τη κατάτμηση κάθε μίας από 7 εικόνες της υπαίθρου, σε πολύ μικρά τμήματα εικονοστοιχείων (pixels) μεγέθους 3x3. Οι 7 κατηγορίες, που περιγράφουν τη φύση του εικονοστοιχείου, είναι: ουρανός, παράθυρο, τσιμέντο, τούβλο, γρασίδι, μονοπάτι και φύλλωμα. Σε κάθε μία κατηγορία αντιστοιχούν 330 πρότυπα, ή αλλιώς τμήματα εικόνας.

Σύνολο δεδομένων Internet Ads

Το σύνολο Internet Ads (Internet advertisements) αποτελείτε από δεδομένα που περιγράφουν εικόνες, οι οποίες περιέχονται σε ιστοσελίδες στο διαδίκτυο. Τα χαρακτηριστικά, τα οποία είναι 1558, περιγράφουν τις διαστάσεις των εικόνων και πιθανές λέξεις ή φράσεις που υπάρχουν στο URL της ιστοσελίδας και της εικόνας καθώς και στο περιεχόμενο της ιστοσελίδας στην περιοχή γύρω από την εικόνα. Αποτελείτε από 3279 πρότυπα και σε ποσοστό 28% οι τιμές των τριών πρώτων χαρακτηριστικών που αφορούν τις διαστάσεις της εικόνας, δεν είναι συμπληρωμένες. Το σύνολο είναι ταξινομημένο σε δύο κατηγορίες αναλόγως με το αν η εικόνα αποτελεί διαφήμιση ή όχι. Τα 458 πρότυπα περιγράφουν διαφημιστικές εικόνες ενώ τα 2821 υπόλοιπα μη διαφημιστικές.

Σύνολο δεδομένων Ionosphere

Το σύνολο δεδομένων Ionosphere προέρχεται από ένα σύστημα ραντάρ στο Goose Bay, του Labrador και αποτελείται από 351 πρότυπα που περιγράφονται

από 34 συνεχή χαρακτηριστικά και κατηγοριοποιούν ελεύθερα ηλεκτρόνια της ιονόσφαιρας σε καλά και κακά (2 κατηγορίες), ανάλογα με το αν σχηματίζουν κάποια δομή ή όχι. Από τα 351 πρότυπα, τα 200 χρησιμοποιούνται για εκπαίδευση του συστήματος, και τα υπόλοιπα 151 για αξιολόγηση, όπως αναφέρεται συνήθως στην βιβλιογραφία.

Σύνολο δεδομένων Iris

Το Iris είναι ίσως το πιο διαδεδομένο σύνολο δεδομένων στη βιβλιογραφία αναγνώρισης προτύπων. Το πρόβλημα έγκειται στην ταξινόμηση του ομώνυμου φυτού σε 3 κατηγορίες και συγκεκριμένα στις Iris Setosa, Iris Virginica και Iris Versicolor. Κάθε κατηγορία περιέχει από 50 πρότυπα (σύνολο 150), ενώ κάθε πρότυπο περιγράφεται από 4 αριθμητικά χαρακτηριστικά του φυτού, που είναι το μήκος και πλάτος του σεπάλου,και το μήκος και πλάτος του πετάλου. Πολύ συχνά στη βιβλιογραφία παρατηρείται το φαινόμενο της "επαναντικατάστασης" (resubstitution) για την αξιολόγηση της ταξινόμησης στο συγκεκριμένο πρόβλημα, που σημαίνει την εξέταση της απόδοσης του συστήματος στα ίδια δεδομένα που εκπαιδεύτηκε, λόγω του μικρού πλήθους προτύπων.

Σύνολο δεδομένων Liver

Το σύνολο Liver (BUPA liver disorders) προέρχεται από την εταιρία BUPA Medical Research Ltd. Κάθε πρότυπο αντιστοιχεί σε μετρήσεις 5 τιμών προερχόμενες από εξετάσεις αίματος και συσχετίζονται με πιθανές ανωμαλίες στο συκώτι από υπερβολική κατανάλωση αλκοόλ. Το έκτο χαρακτηριστικό είναι ένας δείκτης ημερήσιας κατανάλωσης αλκοόλ. Το σύνολο αυτό αποτελείτε από 345 πρότυπα.

Σύνολο δεδομένων Pima Indians

Το σύνολο δεδομένων Pima Indians αποτελείται από 768 πρότυπα που αντιστοιχούν σε άτομα θηλυκού γένους με κληρονομικά χαρακτηριστικά της ινδιάνικης φυλής Pima. Τα 768 αυτά πρότυπα διακρίνονται από 8 βιολογικά αριθμητικά γνωρίσματα και χωρίζονται σε 2 κατηγορίες ανάλογα με το αν το άτομο εμφανίζει συμπτώματα διαβήτη ή όχι.

Σύνολο δεδομένων Sonar

Τα δεδομένα Sonar περιλαμβάνουν 208 πρότυπα σήματος ενός ηχητικού εντοπιστή (sonar) διάστασης 60. Σκοπός είναι ο διαχωρισμός των σημάτων σε δύο κατηγορίες, ανάλογα με το αν το σήμα προσκρούει σε μεταλλικό κυλινδρικό αντικείμενο ή σε έναν σχεδόν κυλινδρικό βράχο.

Σύνολο δεδομένων Vehicle

Το σύνολο Vehicle προέρχεται από το Ινστιτούτο Turing της Γλασκώβης και περιλαμβάνει 846 σιλουέτες οχημάτων υπό διάφορες γωνίες, οι οποίες χωρίζονται σε 4 κατηγορίες: OPEL, SAAB, BUS, VAN. Η κάθε σιλουέτα περιγράφεται από 18 χαρακτηριστικά.

Σύνολο δεδομένων Wine

Το σύνολο Wine (Wine recognition data) περιέχει τα αποτελέσματα από την χημική ανάλυση δειγμάτων κρασιού. Τα δείγμα προέρχονται από τρεις διαφορετικές ποικιλίες κρασιού που αποτελούν και τις κατηγορίες των προτύπων από την ίδια περιοχή της Ιταλίας. Για κάθε δείγμα, από την χημική ανάλυση προέκυψαν 13 ποσοτικές τιμές των διαφορετικών συστατικών που αποτελούν και τα χαρακτηριστικά των δεδομένων εισόδου. Το σύνολο αποτελείτε από 178 πρότυπα, όλα ταξινομημένα σε μία από τρεις κατηγορίες.

Α2. Σύνολα Δεδομένων NIPS

Τα παρακάτω, ειδικά διαμορφωμένα σύνολα δεδομένων, δημιουργήθηκαν για να εξυπηρετήσουν το σκοπό ενός διεθνή διαγωνισμού αξιολόγησης μεθόδων επιλογής χαρακτηριστικών(feature selection challenge-NIPS2003) [27]. Τα σύνολα αυτά είναι διαθέσιμα προς επεξεργασία στην ιστοσελίδα:
http://clopinet.com/isabelle/Projects/NIPS2003

Σύνολο δεδομένων Arcene

Το πρόβλημα Arcene συνίσταται στο διαχωρισμό καρκινογόνων και φυσιολογικών προτύπων από τα δεδομένα ενός φασματογράφου μάζας. Περιλαμβάνει 100 πρότυπα για εκπαίδευση και 100 πρότυπα για έλεγχο. Κάθε πρότυπο περιγράφεται από 10000 χαρακτηριστικά με συνεχείς τιμές, και ταξινομείται σε μία από τις 2 υπάρχουσες κατηγορίες.

Σύνολο δεδομένων Gisette

Το σύνολο Gisette περιέχει δεδομένα που συνιστούν ένα πρόβλημα αναγνώρισης χειρόγραφων χαρακτήρων. Μόνο οι αριθμητικοί χαρακτήρες 4 και 9 περιλαμβάνονται στο πρόβλημα διότι λόγω της ομοιότητάς τους συνιστούν ιδιαίτερη περίπτωση. Το κάθε πρότυπο αποτελείτε 5000 χαρακτηριστικά, από τα οποία τα 2500 είναι πραγματικά και τα υπόλοιπα αποτελούν τεχνητό θόρυβο. Συνολικά, 6750 πρότυπα απαρτίζουν το σύνολο αυτό.

□

Βιβλιογραφία

[1] . Datasets used for classification: comparison of results, Nicholaus Copernicus University, Department of Informatics, Torun, Poland. http://www.phys.uni.torun.pl/kmk/projects/datasets.html.

[2] Abraham, A., and Nath, B. Hybrid intelligent systems design - a review of a decade of research. Technical report (5/2000), Gippsland School of Computing and Information Technology, Monash University, Australia, 5 2000.

[3] Aha, D., Kibler, D., and Albert, M. Instance-based learning algorithms. *Machine Learning 6*, 1 (1991), 37–66.

[4] Akkus, A., Guvenir, H., Andrews, R., Diederich, J., and Tickle, A. K nearest neighbor classification on feature projections. In *Proceedings of the 13th International Conference on Machine Learning (ICML)* (Bari, Italy, 1996), L. Saitta, Ed., vol. 8, Morgan Kaufmann, pp. 12–19.

[5] Andrews, R., Diederich, J., and Tickle, A. Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-Based Systems 8*, 6 (1995), 373–389.

[6] Bengio, Y., Buhmann, J. M., Abu-Mostafa, Y. S., Embrechts, M., and Zurada, J. M. Special issue on neural networks for data mining and knowledge discovery. *IEEE Transactions on Neural Networks 11*, 3 (2000), 545–822.

[7] Bezdek, J., Cios, K., Swiniarski, R., and Pedrycz, W. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers Norwell, MA, USA, 1981.

[8] Blake, C. L., and Merz, C. J. Uci repository of machine learning databases. http://www.ics.uci.edu/mlearn/_MLRepository.html, De-

partment of Information and Computer Science, University of California, Irvine, 1998.

[9] Caridakis, G., Raouzaiou, A., Karpouzis, K., and Kollias, S. Synthesizing gesture expressivity based on real sequences. In *Workshop on Multimodal Corpora: From Multimodal Behaviour Theories to Usable Models, LREC 2006 Conference* (Genoa, Italy, May 24-26 2006).

[10] Caridakis, G., Pateritsas, C., Drosopoulos, A., Stafylopatis, A., and Kollias, S. Probabilistic video-based gesture recognition using self-organizing feature maps. *Submitted to the International Conference on Artificial Neural Networks (ICANN 2007)* (September 9-13 2007).

[11] Cios, K. J., Pedrycz, W., and Swiniarski, R. W. *Data Mining Methods for Knowledge Discovery.* Kluwer, 1998.

[12] Cost, S., and Salzberg, S. A weighted nearest neighbor algorithm for learning with symbolic features. *Machine Learning 10*, 1 (1993), 57–78.

[13] Cover, T., and Hart, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory 13*, 1 (Jan 1967), 21–27.

[14] Darbari, A. Rule extraction from trained ann: A survey. Tech. rep., Institute of Artificial intelligence, Dep. of Computer Science, TU Dresden, 2001.

[15] Dasarathy, B. *Nearest neighbor (NN) norms: NN pattern classification techniques.* Los Alamitos: IEEE Computer Society Press, 1990.

[16] Demiroz, G., and Guvenir, H. Classification by voting feature intervals. In *Proceedings of the 9th European Conference on Machine Learning* (Prague, 1997), Lecture Notes in Computer Science, Springer-Verlag London, UK, pp. 85–92.

[17] Drobics, M., Bodenhofer, U., and Winiwarter, W. Data mining using synergies between self-organizing maps and inductive learning of fuzzy rules. In *Proceedings of the Joint 9th IFSA World Congress and 20th NAFIPS Int. Conf* (Vancouver, July 2001), pp. 1780–1785.

[18] Fan, H., and Ramamohanarao, K. A bayesian approach to use emerging patterns for classification. In *Proceedings of the 14th Australasian database conference* (Darlinghurst, Australia, Australia, 2003), Australian Computer Society, Inc., pp. 39–48.

[19] Fang, G., Gao, W., and Zhao, D. Large vocabulary sign language recognition based on fuzzy decision trees. *IEEE Transactions on Systems, Man and Cybernetics, Part A 34*, 3 (May 2004), 305–314.

[20] Fawcett, T. Using rule sets to maximize roc performance. In *Proceedings of the IEEE International Conference on Data Mining (ICDM-2001)* (San Jose, California, USA, 2001).

[21] Fayyad, U., and Irani, K. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence* (1993), pp. 1022–1027.

[22] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. *Advances in knowledge discovery and data mining*. American Association for Artificial Intelligence Menlo Park, CA, USA, 1996.

[23] Frasconi, P., Gori, M., Kurfess, F., and Sperduti, A. Special issue on integration of symbolic and connectionist systems. *Cognitive Systems Research 3*, 2 (2002), 121–270.

[24] Frossyniotis, D., Pateritsas, C., and Stafylopatis, A. A multi-clustering fusion scheme for data partitioning. *International Journal of Neural Systems 15*, 5 (2005), 391–401.

[25] Fu, L. Learning capacity and sample complexity on expert networks. *IEEE Transactions on Neural Networks 7*, 6 (Nov. 1996), 1517–1520.

[26] Giles, C. L., Sun, R., and Zurada, J. M. Special issue on neural networks and hybrid intelligent models. *IEEE Transactions on Neural Networks 9*, 5 (1998), 721–1054.

[27] Gunn, S. Nips feature selection challenge. http://www.nipsfsc.ecs.soton.ac.uk/, 2003.

[28] Guvenir, H., and Sirin, I. Classification by Feature Partitioning. *Machine Learning 23*, 1 (1996), 47–67.

[29] Guvenir, H., and Akkus, A. Weighted K Nearest Neighbor Classification on Feature Projections. In *Proceedings of the 12th International Symposium on Computer and Information Sciences* (1997), pp. 44–51.

[30] Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L. *Feature Extraction, Foundations and Applications*. Springer, 2006.

[31] Hammerton, J., and Sang, E. Combining a self-organising map with memory-based learning. In *Proceedings of the 2001 workshop on Computational Natural Language Learning* (Toulouse, France, 2001), vol. 7, Association for Computational Linguistics Morristown, NJ, USA, pp. 1–6.

[32] Han, J., and Kamber, M. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2000.

[33] Han, S., and Cho, S. Predicting userős movement with a combination of self-organizing map and markov model. In *Proceedings of the International Conference on Artificial Neural Networks, (ICANN 2006)* (Athens,Greece, Sept. 10-14 2006).

[34] Hastie, T., Tibshirani, R., and Friedman, J. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2001.

[35] Hoeppner, F., Klawonn, F., Kruse, R., and Runkler, T. A. *Fuzzy Cluster Analysis-Methods for Image Recognition, Classification, and Data Analysis*. John Wiley & Sons, Chichester, 1999.

[36] Hong, P., Turk, M., and Huang, T. Gesture modeling and recognition using finite state machines. In *Proceedings of the Fourth IEEE International Conference and Gesture Recognition* (Grenoble, France, March 2000).

[37] Iivarinen, J., Kohonen, T., Kangas, J., and Kaski, S. Visualizing the clusters on the self-organizing map. In *Proceedings of the Conference on Artificial Intelligence Research in Finland* (Helsinki, Finland, 1994), vol. 12, pp. 122–126.

[38] Juang, C., and Ku, K. A recurrent fuzzy network for fuzzy temporal sequence processing and gesture recognition. *IEEE Transactions on Systems, Man and Cybernetics, Part B 35*, 4 (Aug 2005), 646–658.

[39] Karpouzis, K., Raouzaiou, A., Drosopoulos, A., Ioannou, S., Balomenos, T., N., T., and Kollias, S. *3D Modeling and Animation: Synthesis and Analysis Techniques for the Human Body*. Idea

Group Publishing, 2004, ch. Facial Expression and Gesture Analysis for Emotionally-Rich Man-Machine Interaction, pp. 175–200.

[40] Kaski, S. *Data Exploration Using Self-Organizing Maps.* PhD thesis, Helsinki University of Technology, 1997.

[41] Kohonen, T. *Self-organizing maps.* Springer-Verlag, 2000.

[42] Kohonen, T., Kaski, S., Lagus, K., Salojarvi, J., Honkela, J., Paatero, V., and Saarela, A. Self organization of a massive document collection. *IEEE Transactions on Neural Networks 11*, 3 (May 2000), 574–585.

[43] Kononenko, I. Naive bayesian classifier and continuous attributes. *Informatica 16*, 1 (1992), 1–8.

[44] Koskela, T., Varsta, M., Heikkonen, J., and Kaski, K. Recurrent som with local linear models in time series prediction. In *Proceedings of the 6th European Symposium on Artificial Neural Networks (ESANN 98)* (Bruges, Belgium, April 22-24 1998), pp. 167–172.

[45] Levenshtein, V. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady 10* (1966), 707–710.

[46] Malone, J., McGarry, K., Wermter, S., and Bowerman, C. Data mining using rule extraction from Kohonen self-organising maps. *Neural Computing & Applications 15*, 1 (2006), 9–17.

[47] Mantyla, V.-M., Mantyjarvi, J., Seppanen, T., and Tuulari, E. Hand gesture recognition of a mobile device user. In *Proceedings of the IEEE International Conference on Multimedia and Expo, (ICME 2000)* (2000), vol. 1, pp. 281–284.

[48] McGarry, K., Wertmer, S., and MacIntyre, J. Hybrid neural systems: from simple coupling to fully integrated neural networks. *Neural Computing Surveys 2* (1999), 62–93.

[49] Mitchell, T. Machine learning and data mining. *Communications of the ACM 42*, 11 (1999), 30–36.

[50] Mitra, S., Pal, S., and Mitra, P. Data mining in soft computing framework: a survey. *IEEE Transactions on Neural Networks 13*, 1 (Jan. 2002), 3–14.

[51] Moreira, M., and Fiesler, E. Neural networks with adaptive learning rate and momentum terms. Tech. Rep. 95-04, IDIAP, Martigny, Switzerland, 1995.

[52] Ong, S., and Ranganath, S. Automatic sign language analysis: a survey and the future beyond lexical meaning. *IEEE Transactions on Pattern Analysis and Machine Intelligence 27*, 6 (Jun 2005), 873–891.

[53] Ozer, I., Tiehan, L., and Wolf, W. Design of a real-time gesture recognition system: High performance through algorithms and software. *IEEE Signal Processing Magazine 22*, 3 (May 2005), 57– 64.

[54] Parzen, E. On estimation of a probability density function and mode. *Ann. Math. Statistics 33* (1962), 1065–1076.

[55] Pateritsas, C., Modes, S., and Stafylopatis, A. Extracting rules from trained self-organizing maps. In *Proceedings of the International Conference Applied Computing* (Salamanca, Spain, 18-20 February 2007), pp. 183–190.

[56] Pateritsas, C., Pertselakis, M., and Stafylopatis, A. A som-based classifier with enhanced structure learning. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, (SMC 2004)* (The Hague, The Netherlands, 10-13 Oct. 2004), vol. 5, pp. 4832–4837.

[57] Pateritsas, C., and Stafylopatis, A. A nearest features classifier using a self-organizing map for memory base evaluation. In *Proceedings of the 16th International Conference on Artificial Neural Networks (ICANN 2006)* (Athens, Greece, September 10-14 2006), vol. 2, pp. 391–400.

[58] Pateritsas, C., and Stafylopatis, A. Independent nearest features memory-based classifier. In *Proceedings of the International Conference on Computational Intelligence for Modelling, Control and Automation, (CIMCA 2005)* (Vienna, Austria, 28-30 Nov. 2005), vol. 2, pp. 781–786.

[59] Pateritsas, C., and Stafylopatis, A. Memory-based classification with dynamic feature selection using self-organizing maps for pattern evaluation. *International Journal on Artificial Intelligence Tools* (To appear. 2007).

[60] Paul, S., and Kumar, S. Subsethood-product fuzzy neural inference system (supfunis). *IEEE Transactions on Neural Networks 13*, 3 (May 2002), 578–599.

[61] Pertselakis, M., Tsapatsoulis, N., Kollias, S., and Stafylopatis, A. An adaptive resource allocating neural fuzzy inference system. In *Proceedings of the IEEE Intelligent Systems Application to Power Systems, (ISAP'03)* (Limnos, Sep 2003).

[62] Platt, J. A resource-allocating network for function interpolation. *Neural Computation 3*, 2 (1991), 213–225.

[63] Rauber, A. Labelsom: on the labeling of self-organizing maps. In *Proceedings of the International Joint Conference on Neural Networks, (IJCNN '99)* (Washington, DC, 10-16 July 1999), vol. 5, pp. 3527–3532.

[64] Robnik-Sikonja, M., and Kononenko, I. Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning 53* (2003), 23–69.

[65] Sharma, S. *Applied multivariate techniques.* John Wiley & Sons, Inc. New York, NY, USA, 1995.

[66] Siedlecki, W., and Sklansky, J. A note on genetic algorithms for large-scale feature selection. *Pattern Recognition Letters 10* (1989), 335–347.

[67] Solodov, M. V., and Svaiter, B. F. *A comparison of rates of convergence of two inexact proximal point algorithms*, vol. 36 of *Applied Optimization.* Kluwer Academic Publishers, 2000, ch. Nonlinear optimization and related topics, pp. 415–427.

[68] Somervuo, P., and Kohonen, T. Self-organizing maps and learning vector quantization for feature sequences. *Neural Processing Letters 10*, 2 (1999), 151–159.

[69] Stanfill, C., and Waltz, D. Toward memory-based reasoning. *Communications of the ACM 29*, 12 (1986), 1213–1228.

[70] Starner, T., Weaver, J., and Pentland, A. Real-time american sign language recognition using desk and wearable computer-based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence 20*, 12 (Dec 1998), 1371–1375.

[71] Sun, R. *Integrating rules and connectionism for robust commonsense reasoning.* Wiley, 1994.

[72] Sun, R., and Alexandre, F. *Connectionist-Symbolic Integration: From Unified to Hybrid Approaches.* Lawrence Erlbaum Associates, 1997.

[73] Tan, X., Chen, S., Zhou, Z.-H., and Zhang, F. Recognizing partially occluded, expression variant faces from single training image per person with som and soft k-nn ensemble. *IEEE Transactions on Neural Networks 16*, 4 (July 2005), 875–886.

[74] Tickle, A., Andrews, R., Golea, M., and Diederich, J. The truth will come to light: directions and challenges in extracting the knowledge embedded within trained artificial neural networks. *IEEE Transactions on Neural Networks 9*, 6 (Nov. 1998), 1057–1068.

[75] Tomek, I. An Experiment with the Edited Nearest-Neighbor Rule. *IEEE Transactions on Systems, Man, and Cybernetics 6*, 6 (1976), 448–452.

[76] Tong, X., Ozturk, P., and Gu, M. Dynamic feature weighting in nearest neighbor classifiers. In *Proceedings of International Conference on Machine Learning and Cybernetics* (2004), vol. 4.

[77] Ultsch, A. *Knowledge extraction from self-organizing neural networks.* Springer Verlag, Berlin, 1993, ch. Information and Classification, pp. 301–306.

[78] Ultsch, A., and Korus, D. Integration of neural networks with knowledge-based systems. In *Proceedings of the IEEE International Conference on Neural Networks* (Perth, Australia, 27 Nov.-1 Dec. 1995), vol. 4, pp. 1828–1833.

[79] Vesanto, J., Himberg, J., Alhoniemi, E., and Parhankangas, J. Som toolbox for matlab 5. Tech. Rep. A57, Helsinki University of Technology, Apr. 2000.

[80] Vesanto, J., and Alhoniemi, E. Clustering of the self-organizing map. *IEEE Transactions on Neural Networks 11*, 3 (May 2000), 586–600.

[81] Vesanto, J., and Hollmen, J. *Innovations in Intelligent Systems: Design, Management and Applications, Studies in Fuzziness and Soft Computing.* Springer (Physica) Verlag, 2003, ch. An automated report generation tool for the data understanding phase.

[82] Wermter, S., and Sun, R. *Hybrid Neural Systems.* Springer, 2000, ch. An Overview of Hybrid Neural Systems, pp. 1–13.

[83] Wettschereck, D., and Aha, D. Weighting features. In *Proceedings of the First International Conference on Case-Based Reasoning* (Lisbon, Portugal, 1995), Springer-Verlag, pp. 347–358.

[84] Wilson, D. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics 2*, 3 (1972), 408–421.

[85] Wilson, D., and Martinez, T. Improved Heterogeneous Distance Functions. *Journal of Artificial Intelligence Research 6* (1997), 1–34.

[86] Wilson, A., and Bobick, A. Parametric hidden markov models for gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence 21*, 9 (1999), 884–900.

[87] Wilson, D., and Martinez, T. Reduction Techniques for Instance-Based Learning Algorithms. *Machine Learning 38*, 3 (2000), 257–286.

[88] Wong, S., and Cipolla, R. Continuous gesture recognition using a sparse bayesian classifier. In *Proceedings of the 18th international Conference on Pattern Recognition* (Washington, DC, USA, 2006), vol. 1, pp. 1084–1087.

[89] Wu, Y., and Huang, T. Hand modeling, analysis and recognition. *IEEE Signal Processing Magazine 18*, 3 (May 2001), 51–60.

[90] Yang, J., and Honavar, V. Feature subset selection using a genetic algorithm. *IEEE Intelligent Systems 13* (1998), 44–49.

[91] Yang, Y., and Webb, G. Proportional k-interval discretization for naive-Bayes classifiers (ECML). In *Proceedings of the 12th European Conference on Machine Learning* (Freiburg, Germany, 2001), Springer, pp. 564–575.

[92] Yang, M., Ahuja, N., and Tabb, M. Extraction of 2d motion trajectories and its application to hand gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence 24*, 8 (Aug 2002), 1061–1074.

Κατάλογος δημοσιεύσεων του συγγραφέα

ΠΕΡΙΟΔΙΚΑ

[1] Frossyniotis, D., Pateritsas, C., and Stafylopatis, A. A multiclustering fusion scheme for data partitioning. International Journal of Neural Systems 15, 5 (2005), 391-401.

[2] Pateritsas, C., and Stafylopatis, A. Memory-based classification with dynamic feature selection using self-organizing maps for pattern evaluation. International Journal on Artificial Intelligence Tools. To appear (2007).

ΣΥΝΕΔΡΙΑ

[3] Pateritsas, C., Pertselakis, M., and Stafylopatis, A. A Som-based classifier with enhanced structure learning. In Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, (SMC 2004) (The Hague, The Netherlands, 10-13 Oct. 2004), vol. 5, pp. 4832-4837.

[4] Pateritsas, C., and Stafylopatis, A. Independent nearest features memory-based classifier. In Proceedings of the International Conference on Computational Intelligence for Modelling, Control and Automation, (CIMCA 2005) (Vienna, Austria, 28-30 Nov. 2005), vol. 2, pp. 781-786.

[5] Pateritsas, C., and Stafylopatis, A. A nearest features classifier using a self-organizing map for memory base evaluation. In Proceedings of the 16th International Conference on Artificial Neural Networks (ICANN 2006) (Athens, Greece, September 10-14 2006), vol. 2, pp. 391-400.

[6] Pateritsas, C., Modes, S., and Stafylopatis, A. Extracting rules from trained self-organizing maps. In Proceedings of the International Conference Applied Computing (Salamanca, Spain, 18-20 February 2007), pp. 183-190.

[**7**] Caridakis, G., Pateritsas, C., Drosopoulos, A., Stafylopatis, A., and Kollias, S. Probabilistic video-based gesture recognition using self-organizing feature maps. Submitted to the International Conference on Artificial Neural Networks (ICANN 2007) (Porto, Portugal, September 9-13 2007).

□

Βιογραφικό Σημείωμα

Ο Χρήστος Πατερίτσας γεννήθηκε στην Αθήνα τον Δεκέμβριο του 1976. Αποφοίτησε από το 3ο Γενικό Λύκειο Καλλιθέας με βαθμό 17 και 6/10, ενώ τον Οκτώβριο του 2001 απέκτησε το δίπλωμα του Ηλεκτρολόγου Μηχανικού και Μηχανικού Υπολογιστών με βαθμό «Λίαν Καλώς» (7.21) από την σχολή των Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου. Τον Νοέμβριο του 2001 έγινε δεκτός ως υποψήφιος διδάκτορας στην ίδια σχολή με ερευνητικό θέμα στην περιοχή της Υπολογιστικής Νοημοσύνης και των Ευφυών Συστημάτων. Από τον Δεκέμβριο του 2003 είναι υπότροφος του προγράμματος «ΗΡΑΚΛΕΙΤΟΣ: Υποτροφίες έρευνας με προτεραιότητα στη βασική έρευνα» και η διδακτορική διατριβή του αποτελεί υποέργο του προγράμματος αυτού. Διαθέτει πολλές δημοσιεύσεις σε διεθνή συνέδρια και περιοδικά.

Επίσης, προσέφερε επικουρικό εργαστηριακό έργο στο πλαίσιο του μαθήματος "Νευρωνικά Δίκτυα και Ευφυή Υπολογιστικά Συστήματα" (9ο εξάμηνο), με την παρακολούθηση και εργαστηριακή υποστήριξη σπουδαστών κατά την εκπόνηση εργασιών. Ακόμη, συμμετείχε ενεργά στην επίβλεψη διπλωματικών εργασιών.

Εθελοντικοί Επιστημονικοί Ρόλοι

2006	Κριτής άρθρων περιοδικού "IEEE Transactions on Knowledge and Data Engineering"
2006	Κριτής άρθρων συνεδρίου "International Conference on Artificial Neural Networks (ICANN 2006)"

Συμμετοχή σε Ερευνητικά Προγράμματα

2003	«Ελληνικό Πρόγραμμα Δράσεων για την Ασφάλεια Δικτύων και Πληροφοριών, NETIS 2003»

□